Lexical Semantics

(Following slides are modified from Prof. Claire Cardie's slides.)

Introduction to lexical semantics

- Lexical semantics is the study of
  - the systematic meaning-related connections among words and
  - the internal meaning-related structure of each word
- <u>Lexeme</u>
  - an individual entry in the lexicon
  - a pairing of a particular orthographic and phonological form with some form of symbolic meaning representation
- <u>Sense</u>: the lexeme's meaning component
- <u>Lexicon</u>: a finite list of lexemes

Dictionary entries

- right *adj.*
- left *adj.*
- red *n.*
- blood *n.*

Dictionary entries

- right *adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer.
- left *adj.* located nearer to this side of the body than the right.
- red *n.*
- blood *n.*

Dictionary entries

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

Lexical semantic relations: **Homonymy**

- **Homonyms**: *words that have the same form and unrelated meanings*
  - The ***bank***^{1} had been offering 8 billion pounds in 91-day bills.
  - As agriculture burgeons on the east ***bank***^{2}, the river will shrink even more.
- **Homophones**: distinct lexemes with a shared pronunciation
  - E.g. *would* and *wood*, *see* and *sea.*
- **Homographs**: identical orthographic forms, different pronunciations, and unrelated meanings
  - The fisherman was fly-casting for **bass** rather than trout.
  - I am looking for headphones with amazing **bass**.

Lexical semantic relations: **Polysemy**

- Polysemy: the phenomenon of multiple *related* meanings within a single lexeme
    - bank: financial institution as corporation
    - bank: a building housing such an institution
  - **Homonyms** (disconnected meanings)
    - bank: financial institution
    - bank: sloping land next to a river

- Distinguishing homonymy from polysemy is not always easy. Decision is based on:
  - Etymology: history of the lexemes in question
  - Intuition of native speakers

Lexical semantic relations: **Synonymy**

- Lexemes with the same meaning
- Invoke the notion of **substitutability**
  - Two lexemes will be considered synonyms if they can be substituted for one another in a sentence without changing the meaning or acceptability of the sentence
  - How *big* is that plane?
  - Would I be flying on a *large* or small plane?
  - Miss Nelson, for instance, became a kind of *big* sister to Mrs. Van Tassel's son, Benjamin.
  - We frustrate 'em and frustrate 'em, and pretty soon they make a *big* mistake.

Word sense disambiguation (WSD)

- Given a *fixed* set of senses associated with a lexical item, determine which of them applies to a particular instance of the lexical item

- Fundamental question to many NLP applications.
  - Spelling correction
  - Speech recognition
  - Text-to-speech
  - Information retrieval

WordNet

(Following slides are modified from Prof. Claire Cardie's slides.)

WordNet

- Handcrafted database of lexical relations
- Separate databases: nouns; verbs; adjectives and adverbs
- Each database is a set of lexical entries (according to unique orthographic forms)
  - Set of senses associated with each entry

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

WordNet

- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- Try WordNet online at
- http://wordnetweb.princeton.edu/perl/webwn
- How many different meanings for “eat”?
- How many different meanings for “dog”?

Sample entry

The noun "bass" has 8 senses in WordNet.

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

WordNet **Synset**

- Synset == Synonym Set
- Synset is defined by a set of words

- Each synset represents a different "sense" of a word
  - Consider synset == sense

- Which would be bigger?

  # of unique words

  V.S

  # of unique synsets

Statistics

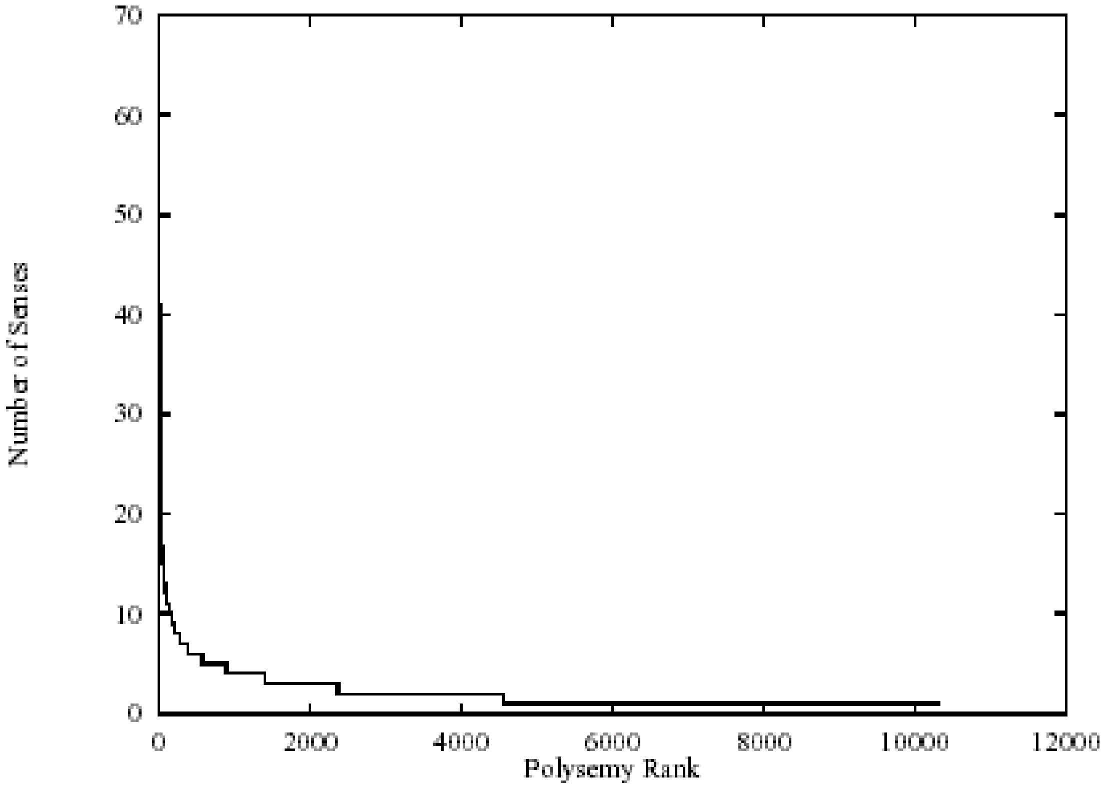
POS	Unique Strings	Synsets	Total word+sense pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adj	21479	18156	30002
Adv	4481	3621	5580
Totals	*155287*	*11765*	*206941*

More WordNet Statistics

Part-of-speech	Avg Polysemy	Avg Polysemy w/o monosemous words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.40	2.71
Adverb	1.25	2.50

Distribution of senses

- Zipf distribution of senses



WordNet relations

- Nouns
- Verbs
- Adjectiv

Relation	Definition	Example
Hypernym	From concepts to superordinates	*breakfast* $\rightarrow$ *meal*
Hyponym	From concepts to subtypes	*meal* $\rightarrow$ *lunch*
Has-Member	From groups to their members	*faculty* $\rightarrow$ *professor*
Member-Of	From members to their groups	*copilot* $\rightarrow$ *crew*
Has-Part	From wholes to parts	*table* $\rightarrow$ *leg*
Part-Of	From parts to wholes	*course* $\rightarrow$ *meal*
Antonym	Opposites	*leader* $\rightarrow$ *follower*

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly* $\rightarrow$ *travel*
Troponym	From events to their subtypes	*walk* $\rightarrow$ *stroll*
Entails	From events to the events they entail	*snore* $\rightarrow$ *sleep*
Antonym	Opposites	*increase* $\Longleftrightarrow$ *decrease*

Relation	Definition	Example
Antonym	Opposite	*heavy* $\Longleftrightarrow$ *light*
Adverb	Opposite	*quickly* $\Longleftrightarrow$ *slowly*

Selectional Preference

Selectional Restrictions & Selectional Preferences

- I want to eat someplace that's close to school.
  - => "eat" is intransitive
- I want to eat Malaysian food.
  - => "eat" is transitive
- "eat" expects its object to be edible.
- What about the subject of "eat"?

Selectional Restrictions & Selectional Preferences

- What are selectional restrictions (or selectional preferences) of…
  - “imagine”
  - “diagonalize”
  - “odorless”

- Some words have stronger selectional preferences than others. How can we quantify the strength of selectional preferences?

Selectional Preference Strength

- P(c) := the distribution of semantic class ‘c’
- P(c|v) := the distribution of semantic class ‘c’ of the object of the given verb ‘v’

- What does it mean if P(c) = P(c|v) ?
- What does it mean if P(c) is very different from P(c|v) ?

- The difference between distributions can be measured by **Kullback-Leibler divergence (KL divergence)**

$$D(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

Selectional Preference Strength

- Selectional preference of ‘v’

$$\begin{aligned} S_R(v) &:= D(P(c|v)||P(c)) \\ &= \sum_c P(c|v) log \frac{P(c|v)}{P(c)} \end{aligned}$$

- Selectional association of ‘v’ and ‘c’

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) log \frac{P(c|v)}{P(c)}$$

- The difference between distributions can be measured by **Kullback-Leibler divergence (KL divergence)**

$$D(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

Selectional Association

- Selectional association of ‘v’ and ‘c’

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) log \frac{P(c|v)}{P(c)}$$

Verb	Direct Object Semantic Class	Assoc	Direct Object Semantic Class	Assoc
read	WRITING	6.80	ACTIVITY	-.20
write	WRITING	7.26	COMMERCE	0
see	ENTITY	5.79	METHOD	-0.01

Remember Pseudowords for WSD?

- Artificial words created by concatenation of two randomly chosen words
- E.g. “banana” + “door” => “banana-door”

- Pseudowords can generate training and test data for WSD automatically. How?

- Issues with pseudowords?

Pseudowords for Selectional Preference?

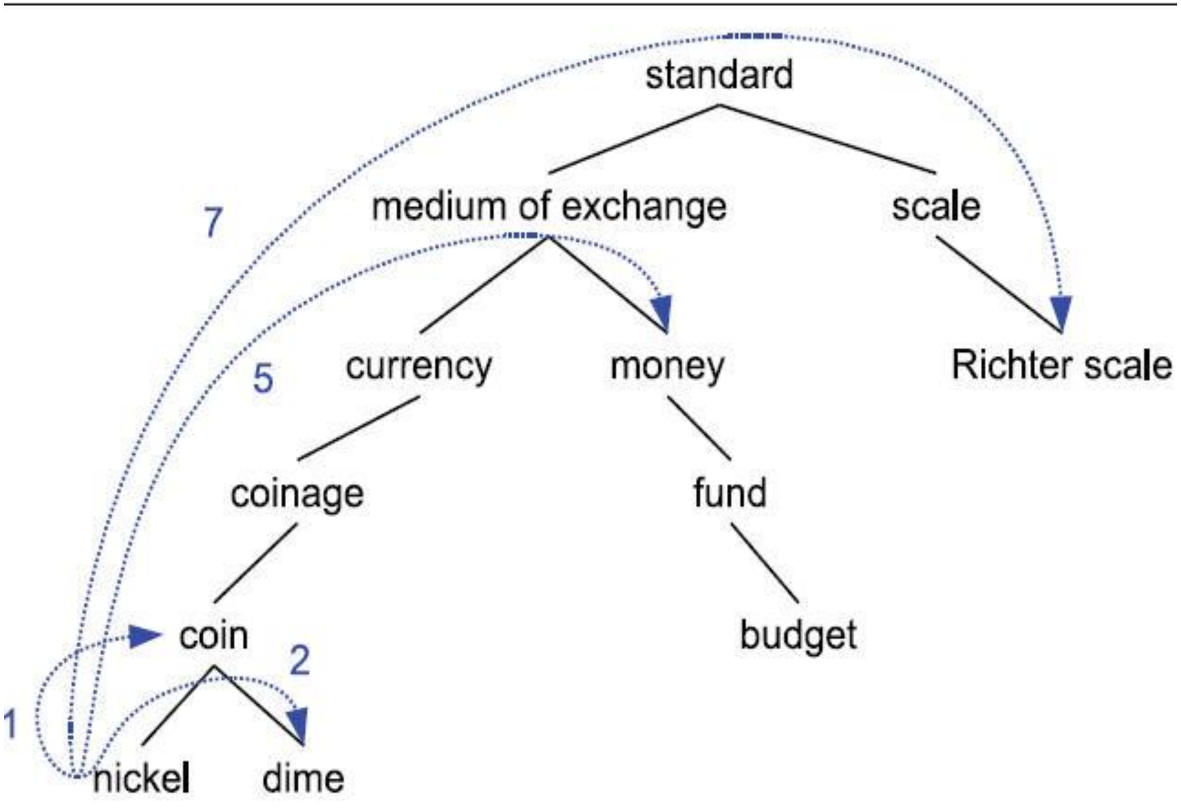
Word Similarity

Word Similarity

- Thesaurus Methods
- Distributional Methods

Word Similarity: Thesaurus Methods

- Path-length based similarity
  - pathlen(nickel, coin) = 1
  - pathlen(nickel, money) = 5



Word Similarity: Thesaurus Methods

- pathlen(x_1, x_2) is the shortest path between x_1 and X_2

- Similarity between two senses --- s_1 and s_2 :

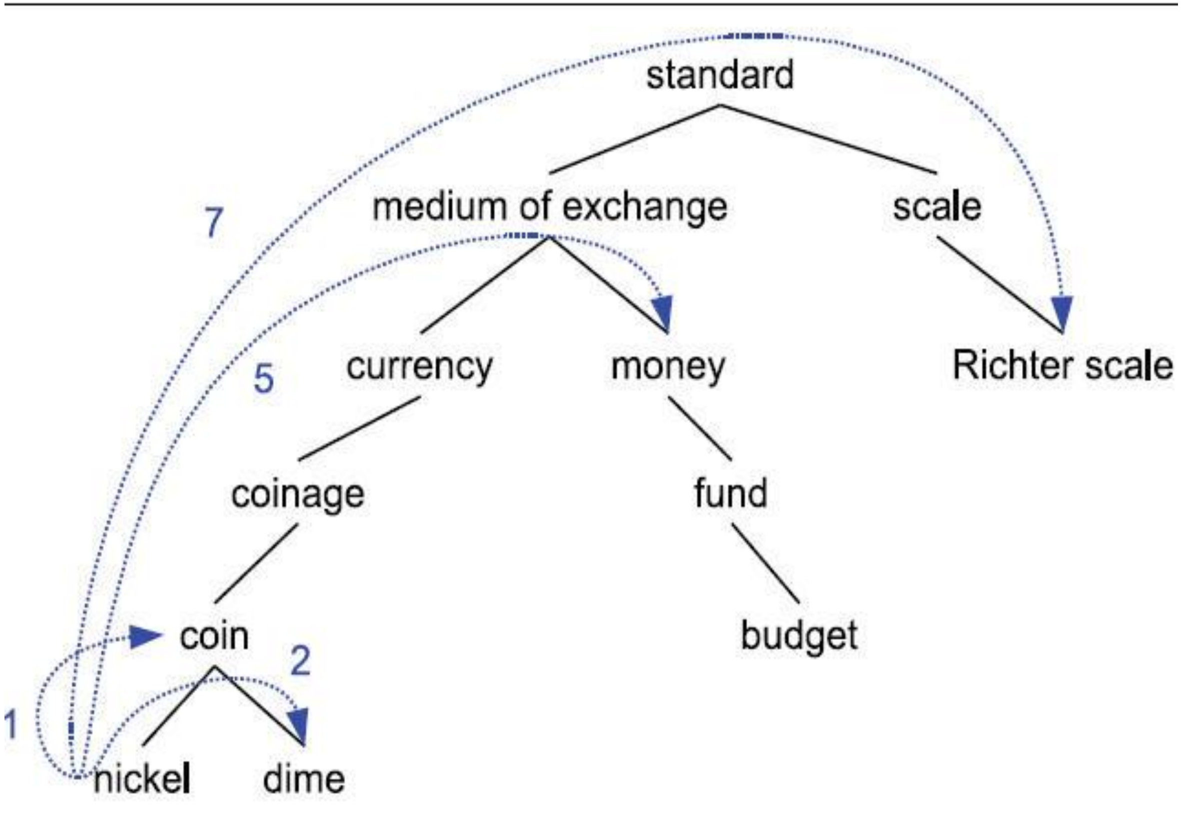
$$\text{sim}_{path}(s_1, s_2) = -\log \text{pathlen}(s_1, s_2)$$

- Similarity between two words --- w_1 and w_2 ?

$$\text{wordsim}(w_1, w_2) = max_{\substack{s_1 \in senses(w_1) \\ s_2 \in senses(w_2)}} \text{sim}(s_1, s_2)$$

Word Similarity: Thesaurus Methods

- Path-length based similarit → ***Problems?***
  - pathlen(nickel, coin) = 1
  - pathlen(nickel, money) = 5



Information-content based word-similarity

- P(c) := the probability that a randomly selected word is an instance of concept 'c'

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

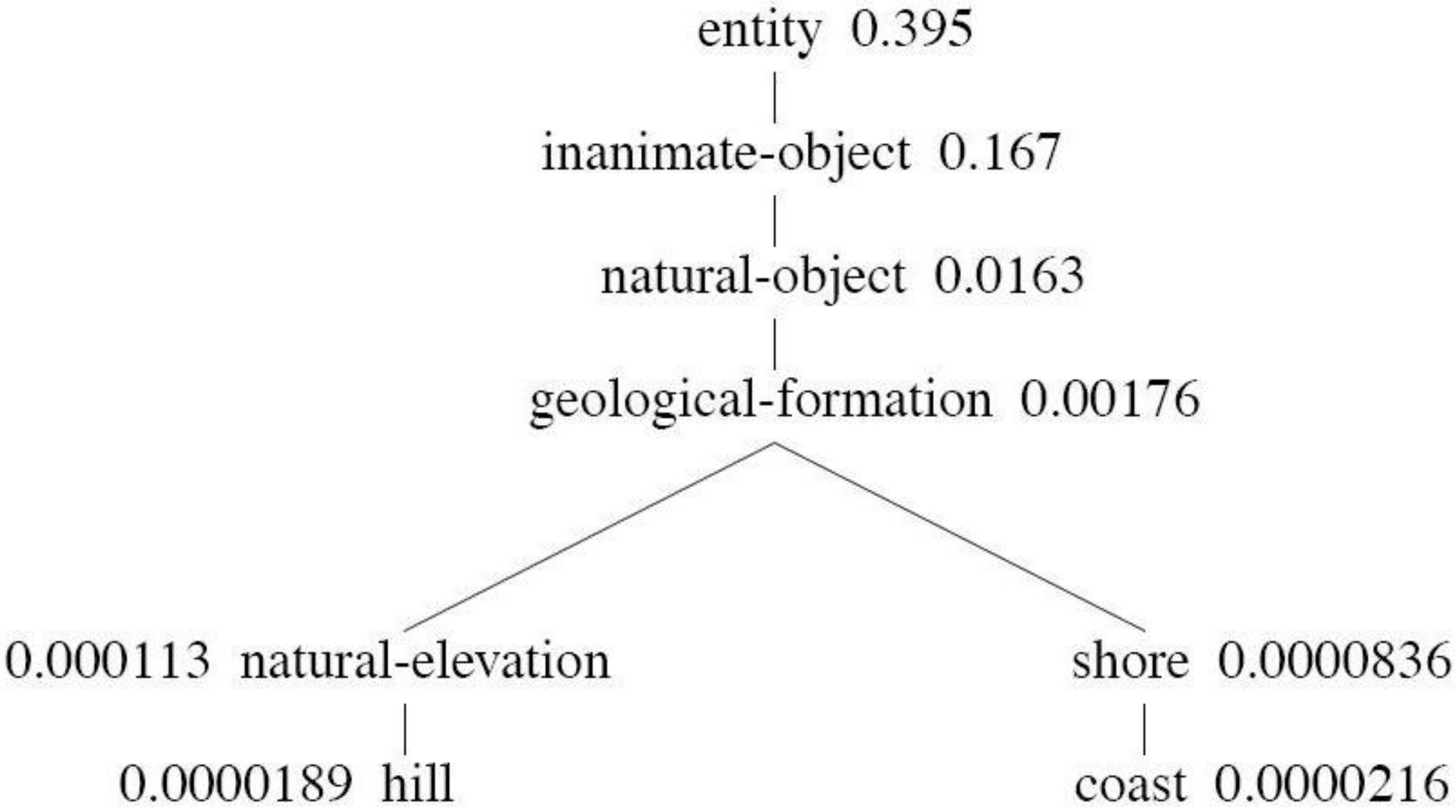
- IC(c) := Information Content

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

- LCS(c_1, c_2) = the lowest common subsumer

$$\mathrm{sim}_{resnik}(c_1, c_2) = -\log P(\mathrm{LCS}(c_1, c_2))$$

Examples of p(c)



Thesaurus-based similarity measures

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

Word Similarity

- Thesaurus Methods

- Distributional Methods

Distributional Word Similarity

- A bottle of tezguino is on the table.
- Tezguino makes you drunk.
- We make tezguino out of corn.

- ***Tezguino, beer, liquor, tequila,*** etc share contextual features such as
  - Occurs before ‘drunk’
  - Occurs after ‘bottle’
  - Is the direct object of ‘likes’

Distributional Word Similarity

- Co-occurrence vectors

	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 Word Similarity

- Co-occurrence vectors <u>with grammatical relations</u>

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

Distributional Word Similarity

	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

Examples of PMI scores

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

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

$$
\begin{aligned}
\text{sim}_{\text{cosine}}(\vec{v},\vec{w}) &= \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}\\
\text{sim}_{\text{Jaccard}}(\vec{v},\vec{w}) &= \frac{\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} \max(v_i,w_i)}\\
\text{sim}_{\text{Dice}}(\vec{v},\vec{w}) &= \frac{2\times\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} (v_i+w_i)}\\
\text{sim}_{\text{JS}}(\vec{v}||\vec{w}) &= D(\vec{v}|\frac{\vec{v}+\vec{w}}{2}) + D(\vec{w}|\frac{\vec{v}+\vec{w}}{2})
\end{aligned}
$$

Distributional Word Similarity

- Problems with Thesaurus-based methods?
  - Some languages lack such resources
  - Thesauruses often lack new words and domain-specific words

- Distributional methods can be used for
  - Automatic thesaurus generation
  - Augmenting existing thesauruses, e.g., WordNet

Vector Space Models for word meaning

(Following slides are modified from Prof. Katrin Erk's slides.)

Geometric interpretation of lists of feature/value pairs

- In cognitive science: representation of a concept through a list of feature/value pairs
- Geometric interpretation:
  - Consider each feature as a dimension
  - Consider each value as the coordinate on that dimension
  - Then a list of feature-value pairs can be viewed as a point in “space”
- Example color represented through dimensions (1) brightness, (2) hue, (3) saturation

Where do the features come from?

- How to construct geometric meaning representations for a large amount of words?
  - Have a lexicographer come up with features (a lot of work)
  - Do an experiment and have subjects list features (a lot of work)

- Is there any way of coming up with features, and feature values, <u>automatically</u>?

Vector spaces: Representing word meaning without a lexicon

- Context words are a good indicator of a word's meaning
- Take a corpus, for example Austen's "Pride and Prejudice"
  Take a word, for example "letter"
- Count how often each other word co-occurs with "letter" in a context window of 10 words on either side

Some co-occurrences:
“letter” in “Pride and Prejudice”

- jane : 12
- when : 14
- by : 15
- which : 16
- him : 16
- with : 16
- elizabeth : 17
- but : 17
- he : 17
- be : 18
- s : 20
- on : 20

- not : 21
- for : 21
- mr : 22
- this : 23
- as : 23
- you : 25
- from : 28
- i : 28
- had : 32
- that : 33
- in : 34

- was : 34
- it : 35
- his : 36
- she : 41
- her : 50
- a : 52
- and : 56
- of : 72
- to : 75
- the : 102

Using context words as features, co-occurrence counts as values

- Count occurrences for multiple words, arrange in a table

context words

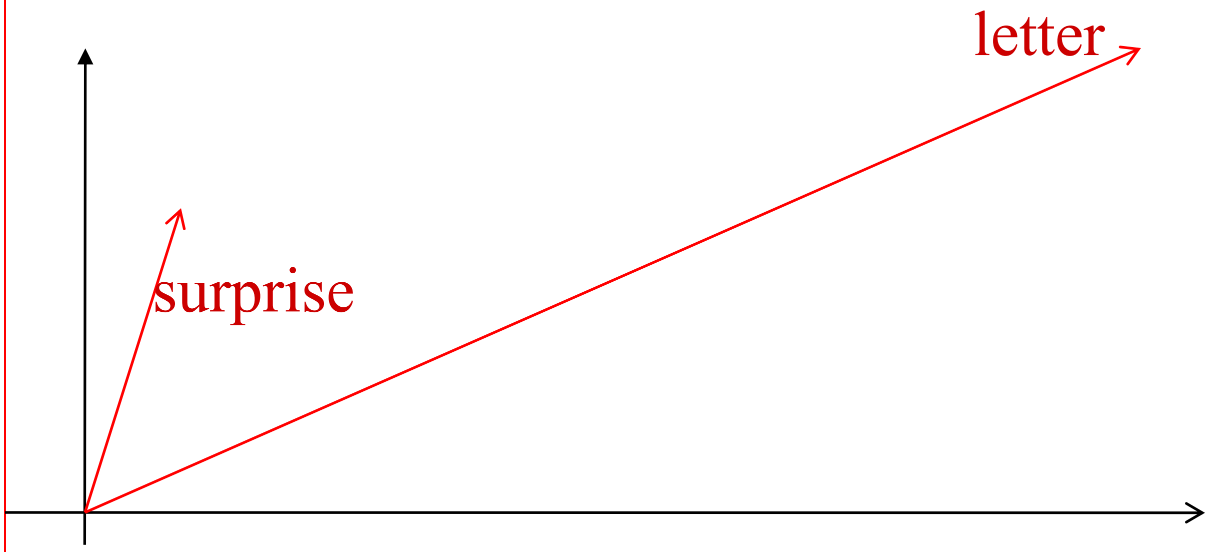
target words

	admirer	all	allow	almost	am	and	angry	...
letter	1	8	1	2	2	56	1	...
surprise	0	7	0	0	4	22	0	...

- For each target word: vector of counts
- Use context words as dimensions
- Use co-occurrence counts as co-ordinates
- For each target word, co-occurrence counts define point in vector space

Vector space representations

- Viewing “letter” and “surprise” as vectors/points in vector space: Similarity between them as distance in space



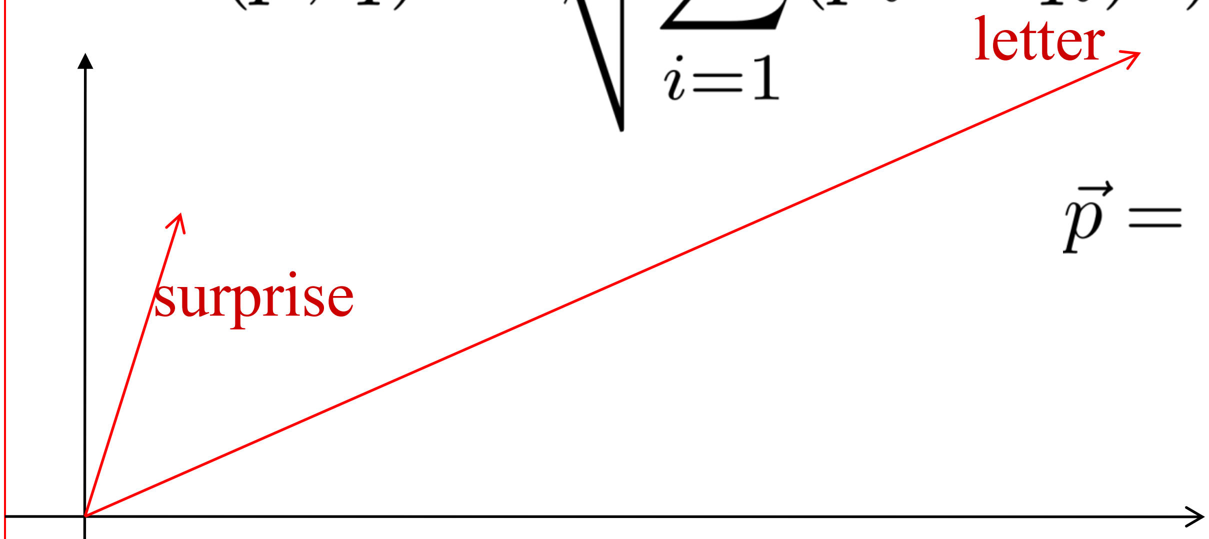
What have we gained?

- Representation of a target word in context space can be computed completely automatically from a large amount of text
- As it turns out, similarity of vectors in context space is a good predictor for semantic similarity
  - Words that occur in similar contexts tend to be similar in meaning
- The dimensions are not meaningful by themselves, in contrast to dimensions like “hue”, “brightness”, “saturation” for color
- Cognitive plausibility of such a representation?

What do we mean by “similarity” of vectors?

<u>Euclidean distance:</u>

$$dist(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2)}$$

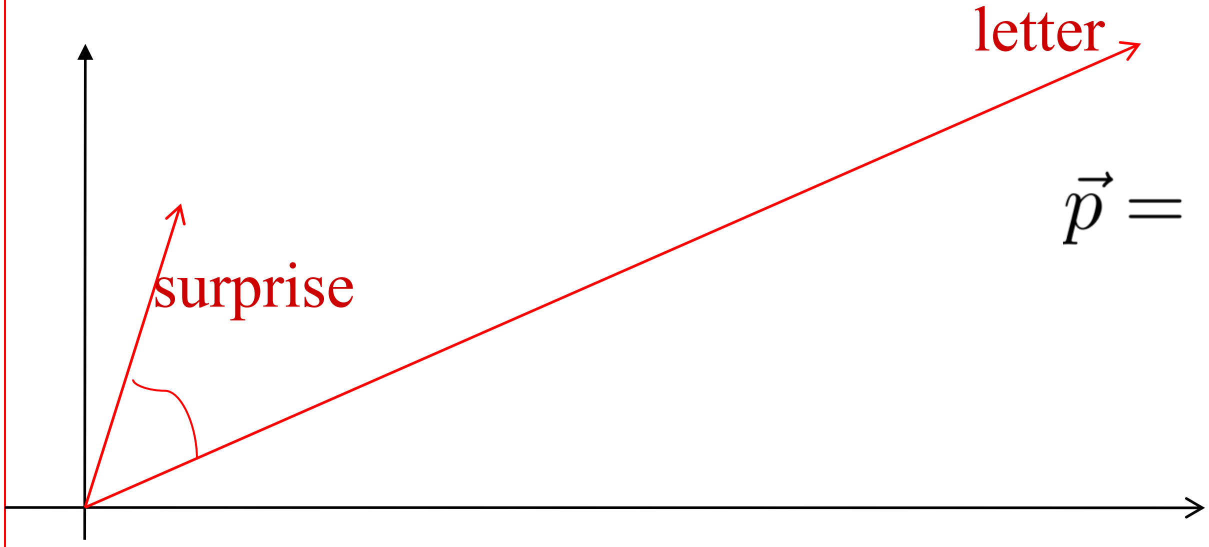


$$\vec{p} = \langle p_1, \ldots, p_n \rangle$$

What do we mean by "similarity" of vectors?

<u>Cosine similarity:</u>

$$cos(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}$$



$$\vec{p} = \langle p_1, \ldots, p_n \rangle$$

Parameters of vector space models

- W. Lowe (2001): "Towards a theory of semantic space"
- A semantic space defined as a tuple
  (A, B, S, M)
- B: base elements. We have seen: context words
- A: mapping from raw co-occurrence counts to something else, for example to correct for frequency effects
  (We shouldn't base all our similarity judgments on the fact that every word co-occurs frequently with 'the')
- S: similarity measure. We have seen: cosine similarity, Euclidean distance
- M: transformation of the whole space to different dimensions (typically, dimensionality reduction)

A variant on B, the base elements

- Term x document matrix:
  - Represent document as vector of weighted terms
  - Represent term as vector of weighted documents

Another variant on B, the base elements

- Dimensions:
  not words in a context window, but dependency paths starting from the target word (Pado & Lapata 07)

	of+pcomp-n	of+mod	in+pcomp-n	in+mod	to+aux	i+subj	he+subj	...
make	124	2426	15810	39	8978	34932	565	...
his	5082	0	0	3682	0	0	83	...

A possibility for A, the transformation of raw counts

- Problem with vectors of raw counts: Distortion through frequency of target word
- Weigh counts:
  - The count on dimension “and” will not be as informative as that on the dimension “angry”
- For example, using Pointwise Mutual Information between target and context word

$$\boxed{PMI(a,b) = \log \frac{P(a,b)}{P(a) \cdot P(b)}}$$

A possibility for M, the transformation of the whole space

- Singular Value Decomposition (SVD): dimensionality reduction
- Latent Semantic Analysis, LSA
  (also called Latent Semantic Indexing, LSI):
  Do SVD on term x document representation
  to induce "latent" dimensions that correspond to topics that a document can be about

  Landauer & Dumais 1997

Using similarity in vector spaces

- Search/information retrieval: Given query and document collection,
  - Use term x document representation:
    Each document is a vector of weighted terms
  - Also represent query as vector of weighted terms
  - Retrieve the documents that are most similar to the query

Using similarity in vector spaces

- To find synonyms:
  - Synonyms tend to have more similar vectors than non-synonyms:
    Synonyms occur in the same contexts
  - But the same holds for antonyms:
    In vector spaces, “good” and “evil” are the same (more or less)
- So: vector spaces can be used to build a thesaurus automatically

Using similarity in vector spaces

- In cognitive science, to predict
  - human judgments on how similar pairs of words are (on a scale of 1-10)
  - “priming”

An automatically extracted thesaurus

- Dekang Lin 1998:
  - For each word, automatically extract similar words
  - vector space representation based on syntactic context of target (dependency parses)
  - similarity measure: based on mutual information ("Lin's measure")
- Large thesaurus, used often in NLP applications

Automatically inducing word senses

- All the models that we have discussed up to now: one vector per word (word type)
- Schütze 1998: one vector per word occurrence (token)
  - She wrote an angry letter to her niece.
  - He sprayed the word in big letters.
  - The newspaper gets 100 letters from readers every day.
- Make token vector by adding up the vectors of all other (content) words in the sentence:

$$\vec{she} + \vec{wrote} + \vec{angry} + \vec{niece}$$

- Cluster token vectors
- Clusters = induced word senses

Summary: vector space models

- Count words/parse tree snippets/documents where the target word occurs
- View context items as dimensions,
target word as vector/point in semantic space
- Distance in semantic space ~
similarity between words
- Uses:
  - Search
  - Inducing ontologies
  - Modeling human judgments of word similarity