LEXICAL SEMANTICS

CS 224N – 2008

Slides largely adapted from ones by Massimo Poesio, Ted Pedersen, Dan Jurafsky, and Jim Martin
Word senses

pike
Lexical information and NL applications

- NL applications often need to know the MEANING of words at least.
- Word meaning is tricky, messy stuff!
- Many word strings express apparently unrelated senses / meanings, even after their POS has been determined.
  - Well-known examples: BANK, SCORE, RIGHT, SET, STOCK
  - Homonymy affects the results of applications such as IR and machine translation.
- The opposite case of different words with the same meaning (SYNONYMY) is also important.
  - NOTEBOOK/LAPTOP
  - E.g., for IR systems (synonym expansion)
- HOMOGRAPHY may affect Speech Synthesis.
An example LEXICAL ENTRY from a machine-readable dictionary: STOCK, from the LDOCE

0100 a supply (of something) for use: a good stock of food
0200 goods for sale: Some of the stock is being taken without being paid for
0300 the thick part of a tree trunk
0400 (a) a piece of wood used as a support or handle, as for a gun or tool (b) the piece which goes across the top of an ANCHOR\(^1\) (1) from side to side
0500 (a) a plant from which CUTTINGS are grown (b) a stem onto which another plant is GRAFTed
0600 a group of animals used for breeding
0700 farm animals usu. cattle; LIVESTOCK
0800 a family line, esp. of the stated character
0900 money lent to a government at a fixed rate of interest
1000 the money (CAPITAL) owned by a company, divided into SHARES
1100 a type of garden flower with a sweet smell
1200 a liquid made from the juices of meat, bones, etc., used in cooking
....
HOMONYMY: Word-strings like STOCK are used to express apparently unrelated senses / meanings, even in contexts in which their part-of-speech has been determined
- Other well-known examples: BANK, RIGHT, SET, SCALE

HOMOGRAPHS: BASS
- The expert angler from Dora, Mo was fly-casting for BASS rather than the traditional trout.
- The curtain rises to the sound of angry dogs baying and ominous BASS chords sounding.
- Problems caused by homography: text to speech

Many spelling errors are caused by HOMOPHONES - distinct lexemes with a single pronunciation
- *Its* vs. *it’s*
- *weather* vs. *whether*
- *their* vs. *there*
POLYSEMY vs HOMONYMY

In cases like BANK, it’s fairly easy to identify two distinct senses (etymology also different). But in other cases, distinctions more questionable

E.g., senses 0100 and 0200 of stock clearly related, like 0600 and 0700, or 0900 and 1000

POLYSEMEOUS WORDS: meanings are related to each other

Cf. human’s foot vs. mountain’s foot

Commonly the result of metaphorical extension

In some cases, syntactic tests may help.

Claim: can conjoin, do ellipsis, etc. over polysemy not homonymy

In general, distinction between HOMONYMY and POLYSEMY not always easy
Two words are **SYNONYMS** if they have the same meaning at least in some contexts

E.g., PRICE and FARE; CHEAP and INEXPENSIVE; LAPTOP and NOTEBOOK; HOME and HOUSE

- I’m looking for a CHEAP FLIGHT / INEXPENSIVE FLIGHT

From Roget’s thesaurus:

- OBLITERATION, erasure, cancellation, deletion

But very few words are truly synonymous in **ALL** contexts:

- I wanna go HOME / ?? I wanna go HOUSE
- The flight was CANCELLED / ?? OBLITERATED / ??? DELETED

Knowing about synonyms may help in IR:

- NOTEBOOK (get LAPTOPs as well)
- CHEAP PRICE (get INEXPENSIVE FARE)
Hyponymy and Hypernymy

- HYPONYMY is the relation between a subclass and a superclass:
  - CAR and VEHICLE
  - DOG and ANIMAL
  - BUNGALOW and HOUSE
- Generally speaking, a hyponymy relation holds between X and Y whenever it is possible to substitute Y for X:
  - That is a X -> That is a Y
  - E.g., That is a CAR -> That is a VEHICLE.
- HYPERNYMY is the opposite relation
- Knowledge about TAXONOMIES useful to classify web pages
  - Eg., Semantic Web. ISA relation of AI
- This information not generally contained explicitly in a traditional or machine-readable dictionary (MRD)
The organization of the lexicon
The organization of the lexicon: Synonymy

"cheap" → CHEAP-LEX-1 → cheap0100

"inexpensive" → INEXP-LEX-3 → inexp0900, inexp1100

WORD-STRINGS LEXEMES SENSES
A free, online, more advanced lexical resource: WordNet

- A lexical database created at Princeton
  - Freely available for research from the Princeton site
  - http://wordnet.princeton.edu/
- Information about a variety of SEMANTICAL RELATIONS
- Three sub-databases (supported by psychological research as early as (Fillenbaum and Jones, 1965))
  - NOUNs
  - VERBS
  - ADJECTIVES and ADVERBS
    - But *no* coverage of closed class parts of speech
- Each database organized around SYNSETS
The noun database

- About 90,000 forms, 116,000 senses
- Relations:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypernym</td>
<td>breakfast -&gt; meal</td>
</tr>
<tr>
<td>hyponym</td>
<td>meal -&gt; lunch</td>
</tr>
<tr>
<td>has-member</td>
<td>faculty -&gt; professor</td>
</tr>
<tr>
<td>member-of</td>
<td>copilot -&gt; crew</td>
</tr>
<tr>
<td>has-Part</td>
<td>table -&gt; leg</td>
</tr>
<tr>
<td>part-of</td>
<td>course -&gt; meal</td>
</tr>
<tr>
<td>antonym</td>
<td>leader -&gt; follower</td>
</tr>
</tbody>
</table>
Synsets

Senses (or `lexicalized concepts') are represented in WordNet by the set of words that can be used in AT LEAST ONE CONTEXT to express that sense / lexicalized concept

- the SYNSET

E.g.,

{chump, fish, fool, gull, mark, patsy, fall guy, sucker, shlemiel, soft touch, mug}

(gloss: person who is gullible and easy to take advantage of)
## Hyponyms

### Sense 1

**robin, redbreast, robin redbreast, Old World robin, Erithacus rubecola** -- (small Old World songbird with a reddish breast)

- => **thrush** -- (songbirds characteristically having brownish upper plumage with a spotted breast)
  - => **oscine, oscine bird** -- (passerine bird having specialized vocal apparatus)
    - => **passerine, passeriform bird** -- (perching birds mostly small and living near the ground with feet having 4 toes arranged to allow for gripping the perch; most are songbirds; hatchlings are helpless)
    - => **bird** -- (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
      - => **vertebrate, craniate** -- (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
        - => **chordate** -- (any animal of the phylum Chordata having a notochord or spinal column)
          - => **animal, animate being, beast, brute, creature, fauna** -- (a living organism characterized by voluntary movement)
            - => **organism, being** -- (a living thing that has (or can develop) the ability to act or function independently)
              - => **living thing, animate thing** -- (a living (or once living) entity)
                - => **object, physical object** --
                  - => **entity, physical thing** --
Meronymy

$ wn \text{beak} - \text{holon}

Holonyms of noun beak

1 of 3 senses of beak

Sense 2

beak, bill, neb, nib

PART OF: bird
The verb database

About 10,000 forms, 20,000 senses

Relations between verb meanings:

<table>
<thead>
<tr>
<th>Hypernym</th>
<th>fly -&gt; travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Troponym</td>
<td>walk -&gt; stroll</td>
</tr>
<tr>
<td>Entails</td>
<td>snore -&gt; sleep</td>
</tr>
<tr>
<td>Antonym</td>
<td>increase -&gt; decrease</td>
</tr>
</tbody>
</table>

V1 ENTAILS V2
when Someone V1 (logically) entails Someone V2
- e.g., snore entails sleep

TROPONYMY
when To do V1 is To do V2 in some manner
- e.g., limp is a troponym of walk
The adjective and adverb database

- About 20,000 adjective forms, 30,000 senses
- 4,000 adverbs, 5600 senses
- Relations:

<table>
<thead>
<tr>
<th>Antonym (adjective)</th>
<th>heavy &lt;-&gt; light</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym (adverb)</td>
<td>quickly &lt;-&gt; slowly</td>
</tr>
</tbody>
</table>
How to use

- Online: [http://wordnet.princeton.edu/perl/webwn](http://wordnet.princeton.edu/perl/webwn)
- Download (various APIs; some archaic)
WORD SENSE DISAMBIGUATION
Identifying the sense of a word in its context

The task of Word Sense Disambiguation is to determine which of various senses of a word are invoked in context:

- the seed companies cut off the tassels of each plant, making it male sterile
- Nissan's Tennessee manufacturing plant beat back a United Auto Workers organizing effort with aggressive tactics

This is generally viewed as a categorization/tagging task

- So, similar task to that of POS tagging
- But this is a simplification!
- Less agreement on what the senses are, so the UPPER BOUND is lower

Word sense discrimination is the problem of dividing the usages of a word into different meanings, without regard to any particular existing sense inventory. Involves unsupervised techniques.

Clear potential uses include Machine Translation, Information Retrieval, Question Answering, Knowledge Acquisition, even Parsing.
- Though in practice the implementation path hasn’t always been clear
Early Days of WSD

- Noted as problem for Machine Translation (Weaver, 1949)
  - A word can often only be translated if you know the specific sense intended (A bill in English could be a pico or a cuenta in Spanish)
- Bar-Hillel (1960) posed the following problem:
  - Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.
  - Is “pen” a writing instrument or an enclosure where children play?
  - ...declared it unsolvable, and left the field of MT (!):
    - “Assume, for simplicity’s sake, that pen in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word pen in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this ‘automatically’.” (1960, p. 159)
"Let me state rather dogmatically that there exists at this moment no method of reducing the polysemy of the, say, twenty words of an average Russian sentence in a scientific article below a remainder of, I would estimate, at least five or six words with multiple English renderings, which would not seriously endanger the quality of the machine output. Many tend to believe that by reducing the number of initially possible renderings of a twenty word Russian sentence from a few tens of thousands (which is the approximate number resulting from the assumption that each of the twenty Russian words has two renderings on the average, while seven or eight of them have only one rendering) to some eighty (which would be the number of renderings on the assumption that sixteen words are uniquely rendered and four have three renderings apiece, forgetting now about all the other aspects such as change of word order, etc.) the main bulk of this kind of work has been achieved, the remainder requiring only some slight additional effort" (Bar-Hillel, 1960, p. 163).
Most early work used semantic networks, frames, logical reasoning, or "expert system" methods for disambiguation based on contexts (e.g., Small 1980, Hirst 1988).

The problem got quite out of hand:
- The word expert for 'throw' is "currently six pages long, but should be ten times that size" (Small and Rieger 1982)

Supervised machine learning sense disambiguation through use of context is frequently extremely successful -- and is a straightforward classification problem.

However, it requires extensive annotated training data.

Much recent work focuses on minimizing need for annotation.
``You shall know a word by the company it keeps''
-- Firth

“You say: the point isn't the word, but its meaning, and you think of the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow that you can buy with it. (But contrast: money, and its use.)"
Wittgenstein, *Philosophical Investigations*

For a large class of cases---though not for all---in which we employ the word `meaning' it can be defined thus: the meaning of a word is its use in the language.''
Wittgenstein, *Philosophical Investigations*
Corpora used for word sense disambiguation work

- **Sense Annotated** (Difficult and expensive to build)
  - Semcor (200,000 words from Brown)
  - DSO (192,000 semantically annotated occurrences of 121 nouns and 70 verbs),
  - Classic words: *interest, line, ...*
  - Training data for Senseval competitions (lexical samples and running text)

- **Non Annotated** (Available in large quantity)
  - Brown, newswire, Web
.....
<wf cmd="ignore" pos="IN">in</wf>
<wf cmd="done" pos="NN" lemma="fig" wnsn="1" lexsn="1:10:00::">fig.</wf>
<wf cmd="done" pos="NN" lemma="6" wnsn="1" lexsn="1:23:00::">6</wf>
<punc>)</punc>
<wf cmd="done" pos="VBP" ot="notag">are</wf>
<wf cmd="done" pos="VB" lemma="slip" wnsn="3" lexsn="2:38:00::">slipped</wf>
<wf cmd="ignore" pos="IN">into</wf>
<wf cmd="done" pos="NN" lemma="place" wnsn="9" lexsn="1:15:05::">place</wf>
<wf cmd="ignore" pos="IN">across</wf>
<wf cmd="ignore" pos="DT">the</wf>
<wf cmd="done" pos="NN" lemma="roof" wnsn="1" lexsn="1:06:00::">roof</wf>
<wf cmd="done" pos="NN" lemma="beam" wnsn="2" lexsn="1:06:00::">beams</wf>
<punc>,</punc>
Dictionary-based approaches

Lesk (1986):
1. Retrieve from MRD all sense definitions of the word to be disambiguated
2. Compare with sense definitions of words in context
3. Choose sense with most overlap

Example:
- PINE
  1. kinds of evergreen tree with needle-shaped leaves
  2. waste away through sorrow or illness
- CONE
  1. solid body which narrows to a point
  2. something of this shape whether solid or hollow
  3. fruit of certain evergreen trees

Disambiguate: **PINE CONE**
Frequency-based word-sense disambiguation

If you have a corpus in which each word is annotated with its sense, you can collect unigram statistics (count the number of times each sense occurs in the corpus)

- $P(\text{SENSE})$
- $P(\text{SENSE}|\text{WORD})$

E.g., if you have

- 5845 uses of the word bridge,
- 5641 cases in which it is tagged with the sense STRUCTURE
- 194 instances with the sense DENTAL-DEVICE

Frequency-based WSD can get about 60-70% correct!

- The WordNet first sense heuristic is good!

To improve upon these results, need context
One type of contextual information is the information about the type of arguments that a verb takes – its SELECTIONAL RESTRICTIONS:

- AGENT EAT FOOD-STUFF
- AGENT DRIVE VEHICLE

Example:
- Which airlines serve DENVER?
- Which airlines serve BREAKFAST?

Limitations:
- In his two championship trials, Mr. Kulkarni ATE GLASS on an empty stomach, accompanied only by water and tea.
- But if fell apart in 1931, perhaps because people realized that you can’t EAT GOLD for lunch if you’re hungry

Resnik (1998): 44% with these methods
Context in general

- But it’s not just classic selectional restrictions that are useful context
  - Often simply knowing the topic is really useful!
Supervised approaches to WSD: the rebirth of Naïve Bayes in CompLing

A Naïve Bayes Classifier chooses the most probable sense for a word given the context:

\[ s = \arg\max P(s_k \mid C) \]

As usual, this can be expressed as:

\[ s = \arg\max \frac{P(C \mid s_k)P(s_k)}{P(C)} \]

The “NAÏVE” ASSUMPTION: all the features are independent

\[ P(C \mid s_k) \approx \prod_{j=1}^{n} P(v_j \mid s_k) \]
An example of use of Naïve Bayes classifiers: Gale et al (1992)

- Used this method to disambiguated word senses using an ALIGNED CORPUS (Hansard) to get the word senses

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>Sense</th>
<th>Number of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>duty</td>
<td>droit</td>
<td>tax obligation</td>
<td>1114 691</td>
</tr>
<tr>
<td></td>
<td>devoir</td>
<td></td>
<td></td>
</tr>
<tr>
<td>drug</td>
<td>medicament</td>
<td>medical illicit</td>
<td>2292 855</td>
</tr>
<tr>
<td></td>
<td>drogue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>land</td>
<td>terre</td>
<td>property country</td>
<td>1022 386</td>
</tr>
<tr>
<td></td>
<td>pays</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Gale et al: words as contextual clues

- Gale et al view a ‘context’ as a set of words
- Good clues for the different senses of DRUG:
  - Medication: prices, prescription, patent, increase, consumer
  - Illegal substance: abuse, paraphernalia, illicit, cocaine, trafficking
- To determine which interpretation is more likely, extract words (e.g. ABUSE) from context, and use $P(\text{abuse} | \text{medicament})$, $P(\text{abuse} | \text{drogue})$ estimated as smoothed relative frequency
- Gale et al (1992): disambiguation system using this algorithm correct for about 90% of occurrences of six ambiguous nouns in the Hansard corpus:
  - duty, drug, land, language, position, sentence
- BUT THIS WAS FOR TWO CLEARLY DIFFERENT SENSES
Gale, Church, and Yarowsky (1992): EDA

Figure II. Remote context is informative.
The horizontal axis shows the distance of context words from an ambiguous word, while the vertical scale shows the percent correct when using ten context words at the specified distance in doing the disambiguation. The vertical lines show the mean and standard deviation of mean for six disambiguations. With two equiprobable choices, 50 percent represents chance performance. Performance remains above chance for ten word contexts up to ten thousand words away from the ambiguous word.
Gale, Church, and Yarowsky (1992): EDA

Figure III. Wide contexts are useful.
The horizontal axis shows the maximum distance of context words from an ambiguous word, while the vertical scale shows the percent correct when using all context words out to the specified distance in disambiguation. While performance rises very rapidly with the first few words, it clearly continues to improve through about twenty words, and is not worse by fifty words.
Gale, Church, and Yarowsky (1992): EDA

Figure IV. Just a few training examples do surprisingly well. The horizontal axis shows the number of examples used in training while the vertical scale shows the mean percent correct in six disambiguations. The performance increases rapidly for the first few examples, and seems to have reached a maximum by 50 or 60 examples.
Other methods for WSD

- **Supervised:**
  - Brown et al, 1991: using mutual information to combine senses into groups
  - Yarowsky (1992): using a thesaurus and a topic-classified corpus
  - More recently, any machine learning method whose name you know

- **Unsupervised:** sense DISCRIMINATION
  - Schuetze 1996: using EM algorithm based clustering, LSA

- **Mixed**
  - Yarowsky’s 1995 bootstrapping algorithm
    - Quite cool
    - A pioneering example of doing context and content constraining each other. More on this later

- **Principles**
  - One sense per collocation
  - One sense per discourse
  - Broad context vs. collocations: both are useful when used appropriately
Evaluation

Baseline: is the system an improvement?
- Unsupervised: Random, Simple-Lesk
- Supervised: Most Frequent, Lesk-plus-corpus.

Upper bound: agreement between humans?
SENSEVAL

Goals:
- Provide a common framework to compare WSD systems
- Standardise the task (especially evaluation procedures)
- Build and distribute new lexical resources

Web site: http://www.senseval.org/

“There are now many computer programs for automatically determining the sense of a word in context (Word Sense Disambiguation or WSD). The purpose of Senseval is to evaluate the strengths and weaknesses of such programs with respect to different words, different varieties of language, and different languages.” from: http://www.sle.sharp.co.uk/senseval2

ACL-SIGLEX workshop (1997): Yarowsky and Resnik paper
SENSEVAL-I (1998); SENSEVAL-II (Toulouse, 2001)
- Lexical Sample and All Words
SENSEVAL-III (2004); SENSEVAL-IV -> SEMEVAL (2007)
WSD at SENSEVAL-II

Choosing the right sense for a word among those of WordNet

Corton has been involved in the design, manufacture and installation of horse stalls and horse-related equipment like external doors, shutters and accessories.

Sense 1: horse, Equus caballus -- (solid-hoofed herbivorous quadruped domesticated since prehistoric times)
Sense 2: horse -- (a padded gymnastic apparatus on legs)
Sense 3: cavalry, horse cavalry, horse -- (troops trained to fight on horseback: "500 horse led the attack")
Sense 4: sawhorse, horse, sawbuck, buck -- (a framework for holding wood that is being sawed)
Sense 5: knight, horse -- (a chessman in the shape of a horse's head; can move two squares horizontally and one vertically (or vice versa))
Sense 6: heroin, diacetyl morphine, H, horse, junk, scag, shit, smack -- (a morphine derivative)
The art of change-ringing is peculiar to the English and, like most English peculiarities, unintelligible to the rest of the world. -- Dorothy L. Sayers, "The Nine Tailors" ASLACTON, England -- Of all scenes that evoke rural England, this is one of the loveliest: An ancient stone church stands amid the fields, the sound of bells cascading from its tower, calling the faithful to evensong. The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have. [...]
English Lexical Sample

- **Data**: 8699 texts for 73 words
- **Average WN polysemy**: 9.22
- **Training Data**: 8166 (average 118/word)
- **Baseline (commonest)**: 0.47 precision
- **Baseline (Lesk)**: 0.51 precision
LEXICAL ACQUISITION:
Lexical and Distributional notions of meaning similarity
How can we work out how similar in meaning words are?

We could use anything in the thesaurus

- Meronymy
- Glosses
- Example sentences

In practice

- By “thesaurus-based” we usually just mean
  - Using the is-a/subsumption/hypernym hierarchy

Word similarity versus word relatedness

- Similar words are near-synonyms
- Related could be related any way
  - Car, gasoline: related, not similar
  - Car, bicycle: similar
Path based similarity

- Two words are similar if nearby in thesaurus hierarchy (i.e. short path between them)
Problem with basic path-based similarity

- Assumes each link represents a uniform distance
- Nickel to money seems closer than nickel to standard
- Instead:
  - Want a metric which lets us represent the cost of each edge independently

- There have been a whole slew of methods that augment thesaurus with notions from a corpus (Resnik, Lin, ...)
  - But I won’t cover them here.
The limits of hand-encoded lexical resources

- Manual construction of lexical resources is very costly
- Because language keeps changing, these resources have to be continuously updated
- Some information (e.g., about frequencies) has to be computed automatically anyway
The coverage problem

Sampson (1989): tested coverage of Oxford ALD (~70,000 entries) looking at a 45,000 subpart of the LOB. About 3% of tokens not listed in dictionary

Examples:

<table>
<thead>
<tr>
<th>Type of problem</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proper noun</td>
<td>Caramello, Chateau-Chalon</td>
</tr>
<tr>
<td>Foreign word</td>
<td>perestroika</td>
</tr>
<tr>
<td>Rare/derived words</td>
<td>reusability</td>
</tr>
<tr>
<td>Code</td>
<td>R101</td>
</tr>
<tr>
<td>Non-standard English</td>
<td>Havin’</td>
</tr>
<tr>
<td>Hyphen omitted</td>
<td>bedclothes</td>
</tr>
<tr>
<td>Technical vocabulary</td>
<td>normoglycaemia</td>
</tr>
</tbody>
</table>
Vector-based lexical semantics

Very old idea in NLE: the meaning of a word can be specified in terms of the values of certain `features' (`COMPONENTIAL SEMANTICS')

- **dog**: ANIMATE=+, EAT=MEAT, SOCIAL=+
- **horse**: ANIMATE=+, EAT=GRASS, SOCIAL=+
- **cat**: ANIMATE=+, EAT=MEAT, SOCIAL=-

Similarity / relatedness: proximity in feature space
Vector-based lexical semantics
General characterization of vector-based semantics

- Vectors as models of concepts
- The CLUSTERING approach to lexical semantics:
  1. Define properties one cares about, and give values to each property (generally, numerical)
  2. Create a vector of length n for each item to be classified
  3. Viewing the n-dimensional vector as a point in n-space, cluster points that are near one another

- What changes between models:
  1. The properties used in the vector
  2. The distance metric used to decide if two points are `close`
  3. The algorithm used to cluster

- For similarity based approaches, skip the 3rd step
Distributional Similarity: Using words as features in a vector-based semantics

The old decompositional semantics approach requires
i. Specifying the features
ii. Characterizing the value of these features for each lexeme
Simpler approach: use as features the WORDS that occur in the proximity of that word / lexical entry
- Intuition: “You shall know a word by the company it keeps.” (J. R. Firth)
- More specifically, you can use as ‘values’ of these features
  - The FREQUENCIES with which these words occur near the words whose meaning we are defining
  - Or perhaps the PROBABILITIES that these words occur next to each other
Using neighboring words to specify the meaning of words

Take, e.g., the following corpus:
1. John ate a banana.
2. John ate an apple.
3. John drove a lorry.

We can extract the following co-occurrence matrix:

<table>
<thead>
<tr>
<th></th>
<th>john</th>
<th>ate</th>
<th>drove</th>
<th>banana</th>
<th>apple</th>
<th>lorry</th>
</tr>
</thead>
<tbody>
<tr>
<td>john</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ate</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>drove</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>banana</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>apple</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lorry</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
To construct vectors $C(w)$ for each word $w$:

1. Scan a text
2. Whenever a word $w$ is encountered, increment all cells of $C(w)$ corresponding to the words $v$ that occur in the vicinity of $w$, typically within a window of fixed size

Differences among methods:

- Size of window
- Weighted or not
- Whether every word in the vocabulary counts as a dimension (including function words such as *the* or *and*) or whether instead only some specially chosen words are used (typically, the $m$ most common content words in the corpus; or perhaps modifiers only). The words chosen as dimensions are often called CONTEXT WORDS
- Whether dimensionality reduction methods are applied
Variant: using modifiers to specify the meaning of words

... The Soviet cosmonaut ... The American astronaut ... The red American car ... The old red truck ... the spacewalking cosmonaut ... the full Moon ...

<table>
<thead>
<tr>
<th></th>
<th>cosmonaut</th>
<th>astronaut</th>
<th>moon</th>
<th>car</th>
<th>truck</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soviet</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>American</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>spacewalking</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>full</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>old</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Measures of semantic similarity

- Euclidean distance:
  \[ d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

- Cosine:
  \[ \cos(\alpha) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \]

- Manhattan (L1) Metric:
  \[ d = \sum_{i=1}^{n} |x_i - y_i| \]
The HAL model (Burgess and Lund, 1995, 1997)

- A 160 million words corpus of articles extracted from all newsgroups containing English dialogue
- Context words: the 70,000 most frequently occurring symbols within the corpus
- Window size: 10 words to the left and the right of the word
- Measure of similarity: cosine

- Frightened: scared, upset, shy, embarrassed, anxious, worried, afraid
- Harmed: abused, forced, treated, discriminated, allowed, attracted, taught
- Beatles: original, band, song, movie, album, songs, lyrics, British
Latent Semantic Analysis (LSA)  
(Landauer et al, 1997)

- Goal: extract relations of expected contextual usage from passages
- Steps:
  1. Build a word / document cooccurrence matrix
  2. `Weight' each cell (e.g., tf.idf)
  3. Perform a DIMENSIONALITY REDUCTION with SVD
- Argued to correlate well with humans on a number of tests
Detecting hyponymy and other relations

- Could we discover new hyponyms, and add them to a taxonomy under the appropriate hypernym?
- Why is this important?
  - “insulin” and “progesterone” are in WordNet 2.1, but “leptin” and “pregnenolone” are not.
  - “combustibility” and “navigability”, but not “affordability”, “reusability”, or “extensibility”.
  - “HTML” and “SGML”, but not “XML” or “XHTML”.
  - “Google” and “Yahoo”, but not “Microsoft” or “IBM”.
- This unknown word problem occurs throughout NLP
Hearst (1992) Approach

- Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.
- What does Gelidium mean? How do you know?

\[ NP_0 \text{ such as } NP_1 \{, NP_2 \ldots ,(\text{and}|\text{or})NP_i \}, i \geq 1 \]

implies the following semantics

\[ \forall NP_i, i \geq 1, \text{hyponym}(NP_i, NP_0) \]

allowing us to infer

\[ \text{hyponym}(\text{Gelidium, red algae}) \]
Hearst's hand-built patterns

| NP{,NP} * {,} (and|or) other NP_H | ... temples, treasuries, and other important civic buildings. |
| NP_H such as {NP,}* (or|and) NP | red algae such as Gelidium |
| such NP_H as {NP,}* (or|and) NP | works by such authors as Herrick, Goldsmith, and Shakespeare |
| NP_H {,} including {NP,}* (or|and) NP | All common-law countries, including Canada and England |
| NP_H {,} especially {NP,}* (or|and) NP | ... most European countries, especially France, England, and Spain |
Recording the Lexico-Syntactic Environment with MINIPAR Syntactic Dependency Paths

MINIPAR: A dependency parser (Lin, 1998)
Example Word Pair: “oxygen / element”
Example Sentence: “Oxygen is the most abundant element on the moon.”

Minipar Parse:

Extracted dependency path:
-N:s:VBE, “be” VBE:pred:N
Each of Hearst’s patterns can be captured by a syntactic dependency path in MINIPAR:

Hearst Pattern

**Y such as X...**

![Diagram of Y such as X...]

\[-N:pcomp-n:Prep, such\_as, such\_as, -Prep:mod:N\]

**Such Y as X...**

![Diagram of Such Y as X...]

\[-N:pcomp-n:Prep, as, as, -Prep:mod:N, (such, PreDet:pre:N)\]}

**X... and other Y**

![Diagram of X... and other Y]

\[(and, U:punc:N), N:conj:N, (other, A:mod:N)\]
Algorithm (Snow, Jurafsky, and Ng 2005)

- Collect noun pairs from corpora
  - (752,311 pairs from 6 million words of newswire)
- Identify each pair as positive or negative example of hypernym-hyponym relationship
  - (14,387 yes, 737,924 no)
- Parse the sentences, extract patterns
  - (69,592 dependency paths occurring in >5 pairs)
- Train a hypernym classifier on these patterns
  - We could interpret each path as a binary classifier
  - Better: **logistic regression** with 69,592 features
    - (actually converted to 974,288 bucketed binary features)
Using Discovered Patterns to Find Novel Hyponym/Hypernym Pairs

Example of a discovered high-precision path:

Learned from cases such as:

“sarcoma / cancer”: ...an uncommon bone cancer called osteogenic sarcoma and to...
“deuterium / atom” ....heavy water rich in the doubly heavy hydrogen atom called deuterium.

May be used to discover new hypernym pairs not in WordNet:

“efflorescence / condition”: ...and a condition called efflorescence are other reasons for...
“neal_inc / company” ...The company, now called O'Neal Inc., was sole distributor of E-Ferol...
“hat_creek_outfit / ranch” ...run a small ranch called the Hat Creek Outfit.
“tardive_dyskinesia / problem”: ... irreversible problem called tardive dyskinesia...
“hiv-1 / aids_virus” ...infected by the AIDS virus, called HIV-1.
“bateau_mouche / attraction” ...local sightseeing attraction called the Bateau Mouche...
“kibbutz_malkiyya / collective_farm” ...an Israeli collective farm called Kibbutz Malkiyya...

But 70,000 patterns are better than one!
Using each pattern/feature as a binary classifier: Hypernym Precision / Recall
There are lots of fun lexical semantic tasks: Logical Metonymy


Additional meaning arises from characterization of an event:

- Mary finished her dinner -->
  - Mary finished eating her dinner
- Mary finished her beer -->
  - Mary finished drinking her beer
  - NOT Mary finished eating her beer
- Mary finished her sweater -->
  - Mary finished knitting her sweater
  - NOT Mary finished eating her sweater

- How can we work out the implicit activities?
Logical metonymy

- Easy cake --> easy stew to make
- Good soup --> good to eat NOT enjoyable to make
- Fast plane --> flies fast NOT fast to construct

There is a default interpretation, but it depends on context:
- All the office personnel took part in the company sports day last week.
- One of the programmers was a good athlete, but the other was struggling to finish the events.
- The fast programmer came first in the 100m.

Some cases seem to lack default metonymic interpretations
- John enjoyed the dictionary
How can you learn them? (Lapata and Lascarides 1999)

- Corpus statistics!
- But cases that fill in the metonymic interpretation (*begin V NP* or *like V NP*) are too rare -- not regularly used
- So just use general facts about verb complements
  - The likelihood of an event is assumed independent of whether it is the complement of another verb.
    - \( P(o|e,v) \approx P(o|e) \)
- Examples learned by model:
  - Begin story --> begin to tell story
  - Begin song --> begin to sing song
  - Begin sandwich --> begin to bite into sandwich
  - Enjoy book --> enjoy reading book

- This doesn’t do context-based interpretation, of course!