# CS114 Lecture 15

## Lexical Semantics

March 19, 2014

Professor Meteer

Thanks for Jurafsky & Martin & Prof. Pustejovksy for slides

# Assignment 3: Superchunks

- Create a new chunker which takes the chunked data and produces better chunks
- Your program should consist of
  - a set of declaratives rules (taking advantage of python data structures)
  - an "interpreter" which is agnostic as to the specific set of rules being applied.
- Use the the dev set from the previous assignment
- Focus your efforts on people, places, and organizations

# Example

- **wsj_0014**

  Norman Ricken, 52 years old and former president and chief operating officer of Toys "R" Us Inc., and Frederick Deane Jr., 63, chairman of Signet Banking Corp., were elected directors of this consumer electronics and appliances retailing chain.

  They succeed Daniel M. Rexinger, retired Circuit City executive vice president, and Robert R. Glauber, U.S. Treasury undersecretary, on the 12-member board.

[ Norman/NNP Ricken/NNP ] ,/, [ 52/CD years/NNS ] old/JJ **and/CC** [ former/JJ president/NN ] **and/CC** [ chief/NN operating/VBG officer/NN ] of/IN [ Toys/NNPS ] ``/`` [ R/NNP ] ''/'' [ Us/NNP ] [ Inc./NNP ] ,/,

**and/CC**

[ Frederick/NNP Deane/NNP Jr./NNP ] ,/, [ 63/CD ] ,/, [ chairman/NN ] of/IN [ Signet/NNP Banking/NNP Corp./NNP ] ,/,

were/VBD elected/VBN [ directors/NNS ] of/IN [ this/DT consumer/NN electronics/NNS ] **and/CC** [ appliances/NNS ] retailing/NN [ chain/NN ] ./

# Three Perspectives on Meaning

1. **Lexical Semantics**
   - The meanings of individual words
2. **Formal Semantics** (or Compositional Semantics or Sentential 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 Conversation** is often lumped together with Discourse

# Outline: Comp Lexical Semantics

- Intro to Lexical Semantics
  - Homonymy, Polysemy, Synonymy
  - Online resources: WordNet
- Computational Lexical Semantics
  - Word Sense Disambiguation
    - Supervised
    - Semi-supervised
  - Word Similarity
    - Thesaurus-based
    - Distributional

# Preliminaries

- What's a word?
  - Definitions we've used: Types, tokens, stems, roots, inflected forms, etc...
  - Lexeme: An entry in a lexicon consisting of a pairing of a form with a single meaning representation
  - Lexicon: A collection of lexemes

# Relationships between word meanings

- What's in a nym?
  - Homonymy
  - Synonymy
  - Antonymy
  - Metonymy
- And a sem?
  - Polysemy
  - Monosemy
- And a nom
  - Hypernomy
    - Hypernym
  - Hyponomy
    - Hyponym
  - Meronomy
  - Holonymy

# Homonymy: Bat and Bass

- Homonymy:
  - Lexemes that share a form
    - Phonological, orthographic or both
  - But have unrelated, distinct meanings
  - Clear example:
    - Bat (wooden stick-like thing) vs
    - Bat (flying scary mammal thing)
    - Or bank (financial institution) versus bank (riverside)
  - Can be homophones, homographs, or both:
    - Homophones:
      - Write and right
      - Piece and peace
    - Homographs
      - Bass
      - Convert

# Homonymy causes problems for NLP applications

- Text-to-Speech
  - Same orthographic form but different phonological form
    - bass vs bass
- Information retrieval
  - Different meanings same orthographic form
    - QUERY: bat care
- Machine Translation
- Speech recognition
  - Why?

# Polysemy

- The **bank** is constructed from red brick
  I withdrew the money from the **bank**
- Are those the same sense?
- Or consider the following WSJ example
  - While some banks furnish sperm only to married women, others are less restrictive
  - Which sense of bank is this?
    - Is it distinct from (homonymous with) the river bank sense?
    - How about the savings bank sense?

# Polysemy

- A single lexeme with multiple related meanings (bank the building, bank the financial institution)
- Most non-rare words have multiple meanings
  - The number of meanings is related to its frequency
  - Verbs tend more to polysemy
  - Distinguishing polysemy from homonymy isn't always easy (or necessary)

# Metaphor and Metonymy

- Specific types of polysemy
- Metaphor:
  - Germany will **pull** Slovenia **out** of its economic slump.
  - I **spent** 2 hours on that homework.
- Metonymy
  - **The White House** announced yesterday.
  - **This chapter** talks about part-of-speech tagging
  - Bank (building) and bank (financial institution)

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

- ATIS examples
  - Which flights serve breakfast?
  - Does America West serve Philadelphia?
- The “zeugma” test:
  - ?Does United serve breakfast and San Jose?

# 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 $H_20$

# 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

# 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 sense is more specific, denoting a subclass of the other
  - *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 |

# Hypernymy 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 it is:
  - http://wordnet.princeton.edu/

# Format of Wordnet Entries

- The noun "bass" has 8 senses in WordNet.
  1. bass[1] - (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 - (the lean flesh of a saltwater fish of the family Serranidae)
  5. freshwater bass , bass - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
  6. bass[6], bass voice[1], basso" - (the lowest adtilt male singing voice)
  7. bass[7] - (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)

- The adjective "bass" has 1 sense in WordNet.

  1. bass , 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] → meal[1] |
| Hyponym | Subordinate | From concepts to subtypes | meal[1] → lunch[1] |
| Member Meronym | Has-Member | From groups to their members | faculty[2] → professor[1] |
| Has-Instance | | From concepts to instances of the concept | composer[1] → Bach[1] |
| Instance | | From instances to their concepts | Austen[1] → author[1] |
| Member Holonvm | Member-Of | From members to their groups | copilot[1] → crew[1] |
| Part Meronym | Has-Part | From wholes to parts | table[2] → leg[3] |
| Part Holonvm | Part-Of | From parts to wholes | course[7] → meal[1] |
| Antonym | | Opposites | leader[1] → follower[1] |

# WordNet Verb Relations

| Relation | Definition | Example |
|---|---|---|
| Hypernym | From events to superordinate events | Fly[9] → travel[5] |
| Troponym | From a verb (event) to a specific manner elaboration of that verb | walk[1] → stroll[1] |
| Entails | From verbs (events) to the verbs (events) they entail | snore[1] → sleep[1] |
| Antonym | Opposites | increase[1] <=> 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
```

# 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.

# Word Sense Disambiguation (WSD)

- Given
  - a word in context,
  - A fixed inventory of potential word senses
- decide which sense of the word this is.
  - English-to-Spanish MT
    - Inventory is set of Spanish translations
  - Speech Synthesis
    - Inventory is homographs with different pronunciations like *bass* and *bow*
  - Automatic indexing of medical articles
    - MeSH (Medical Subject Headings) thesaurus entries

# Two variants of WSD task

- Lexical Sample task
  - Small pre-selected set of target words
  - And inventory of senses for each word
  - We'll use **supervised machine learning**
- All-words task
  - Every word in an entire text
  - A lexicon with senses for each word
  - Sort of like part-of-speech tagging
    - Except each lemma has its own tagset

# Supervised Machine Learning Approaches

- Supervised machine learning approach:
  - a training corpus of words tagged in context with their sense
  - used to train a classifier that can tag words in new text
  - Just as we saw for part-of-speech tagging, statistical MT.
- Summary of what we need:
  - the **tag set** ("sense inventory")
  - the **training corpus**
  - A set of **features** extracted from the training corpus
  - A **classifier**

# Supervised WSD 1: WSD Tags

- What's a tag?
  - A dictionary sense?
- For example, for WordNet an instance of "bass" in a text has 8 possible tags or labels (bass1 through bass8).

# WordNet Bass

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)

# Inventory of sense tags for *bass*

| WordNet Sense | Spanish Translation | Roget Category | Target Word in Context |
|---|---|---|---|
| bass$^4$ | lubina | FISH/INSECT | ... fish as Pacific salmon and striped **bass** and... |
| bass$^4$ | lnbina | FISH/INSECT | ... produce filets of smoked **bass** or sturgeon... |
| bass$^7$ | bajo | MUSIC | ... exciting jazz **bass** player since Ray Brown... |
| bass$^7$ | bajo | MUSIC | .. .play **bass** because he doesn't have to solo... |

# Supervised WSD 2: Get a corpus

- Lexical sample task:
  - *Line-hard-serve* corpus - 4000 examples of each
  - *Interest* corpus - 2369 sense-tagged examples
- All words:
  - **Semantic concordance**: a corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus.
    - SemCor: 234,000 words from Brown Corpus, manually tagged with WordNet senses
    - SENSEVAL-3 competition corpora - 2081 tagged word tokens

# Supervised WSD 3: Extract feature vectors

- Weaver (1955)
  - If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. […]
  - But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word. […]
  - The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?''

# Feature vectors

- A simple representation for each observation (each instance of a target word)
  - Vectors of sets of feature/value pairs
    - I.e. files of comma-separated values
  - These vectors should represent the window of words around the target

# Two kinds of features in the vectors

- **Collocational** features and **bag-of-words** features
  - **Collocational**
    - Features about words at **specific** positions near target word
      - Often limited to just word identity and POS
  - **Bag-of-words**
    - Features about words that occur anywhere in the window (regardless of position)
      - Typically limited to frequency counts

# Examples

- Example text (WSJ)
  - An electric guitar and **bass** player stand off to one side not really part of the scene, just as a sort of nod to gringo expectations perhaps
  - Assume a window of +/- 2 from the target

# Examples

- Example text
  - An electric guitar and **bass** player stand off to one side not really part of the scene, just as a sort of nod to gringo expectations perhaps
  - Assume a window of +/- 2 from the target

# Collocational

- Position-specific information about the words in the window
- guitar and bass player stand
  - [guitar, NN, and, CC, player, NN, stand, VB]
  - $Word_{n-2,}$ $POS_{n-2,}$ $word_{n-1,}$ $POS_{n-1,}$ $Word_{n+1}$ $POS_{n+1}$…
  - In other words, a vector consisting of
  - [position n word, position n part-of-speech…]

# Bag-of-words

- Information about the words that occur within the window.
- First derive a set of terms to place in the vector.
- Then note how often each of those terms occurs in a given window.

# Co-Occurrence Example

- Assume we've settled on a possible vocabulary of 12 words that includes guitar and player but not and and stand

- guitar and bass player stand
  - [0,0,0,1,0,0,0,0,0,1,0,0]
  - Which are the counts of words predefined as e.g.,
  - [fish, fishing, viol, guitar, double, cello…

# Classifiers

- Once we cast the WSD problem as a classification problem, then all sorts of techniques are possible
  - Naïve Bayes (the easiest thing to try first)
  - Decision lists
  - Decision trees
  - Neural nets
  - Support vector machines
  - Nearest neighbor methods…