Lecture 20:
Lexical Semantics:
Word Sense

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center
Part 1: Lexicographic approaches to word meaning
Where we’re at

We have looked at how to represent the meaning of sentences based on the meaning of their words (using predicate logic).

Now we will get back to the question of how to represent the meaning of words (although this won’t be in predicate logic).

We will look at lexical resources (WordNet)

We will consider two different tasks:

— Computing word similarities
— Word sense disambiguation
Different approaches to lexical semantics

Lexicographic tradition (today’s lecture)

- Use lexicons, thesauri, ontologies
- Assume words have discrete word senses:
  \( \text{bank1} = \text{financial institution}; \ \text{bank2} = \text{river bank}, \text{etc.} \)
- May capture explicit relations between word (senses):
  “dog” is a “mammal”, etc.

Distributional tradition (earlier lectures)

- Map words to (sparse) vectors that capture corpus statistics
- Contemporary variant: use neural nets to learn dense vector “embeddings” from very large corpora
  (this is a prerequisite for most neural approaches to NLP)
- This line of work often ignores the fact that words have multiple senses or parts-of-speech
Word senses

What does ‘bank’ mean?

– a financial institution
  *(US banks have raised interest rates)*

– a particular branch of a financial institution
  *(the bank on Green Street closes at 5pm)*

– the bank of a river
  *(In 1927, the bank of the Mississippi flooded)*

– a ‘repository’
  *(I donate blood to a blood bank)*
Lexicon entries

bank¹ /baNGk/  
noun  
1. the land alongside or sloping down to a river or lake: willows lined the riverbank.  
2. a slope, mass, or mound of a particular substance: a bank of clouds | a bank of snow.  
   • an elevation in the seabed or a riverbed; a mudbank or sandbank.  
   • a transverse slope given to a road, railroad, or sports track to enable vehicles or runners to maintain speed around a curve.  
   • the sideways tilt of an aircraft when turning in flight: flying with small amounts of bank.  
3. a set or series of similar things, esp. electrical or electronic devices, grouped together in rows: the DJ had big banks of lights and speakers on either side of his console.  
   • a tier of oars: the early ships had only twenty-five oars in each bank.  
4. the cushion of a pool table: [as adj.] a bank shot.

bank²  
noun  
5. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency: I paid the money straight into my bank.  
   • a stock of something available for use when required: a blood bank | building a bank of test items is the responsibility of teachers.  
   • a place where something may be safely kept: the computer’s memory bank.  
   • (the bank) the store of money or tokens held by the banker in some gambling or board games.  
   • the person holding this store; the banker.  
   • Brit. a site or receptacle where something may be deposited for recycling: a paper bank.
Some terminology

**Word forms:** *runs, ran, running; good, better, best*

Any, possibly inflected, form of a word
(i.e. what we talked about in morphology)

**Lemma** (citation/dictionary form): *run*

A basic word form (e.g. infinitive or singular nominative noun) that is used to represent all forms of the same word.
(i.e. the form you’d search for in a dictionary)

**Lexeme:** *Run(V), Good(A), Bank\(^1\)(N), Bank\(^2\)(N)*

An abstract representation of a word (and all its forms), with a part-of-speech and a set of related word senses.
(Often just written (or referred to) as the lemma, perhaps in a different font)

**Lexicon:**

A (finite) list of lexemes
Trying to make sense of senses

Polysemy:
A lexeme is polysemous if it has different related senses

bank = financial institution or building

Homonyms:
Two lexemes are homonyms if their senses are unrelated, but they happen to have the same spelling and pronunciation

bank = (financial) bank or (river) bank
Relations between senses

**Symmetric** relations:

**Synonyms**: couch/sofa
Two lemmas with the **same** sense

**Antonyms**: cold/hot, rise/fall, in/out
Two lemmas with the **opposite** sense

**Hierarchical** relations:

**Hypernyms** and **hyponyms**: pet/dog
The hyponym *(dog)* is **more specific** than the hypernym *(pet)*

**Holonyms** and **meronyms**: car/wheel
The meronym *(wheel)* is a **part of** the holonym *(car)*
WordNet

Very large **lexical database** of English:

110K nouns, 11K verbs, 22K adjectives, 4.5K adverbs

(WordNets for many other languages exist or are under construction)

**Word senses** grouped into synonym sets ("**synsets**") linked into a **conceptual-semantic hierarchy**

81K noun synsets, 13K verb synsets, 19K adj. synsets, 3.5K adv synsets

Avg. # of senses: 1.23 nouns, 2.16 verbs, 1.41 adj, 1.24 adverbs

**Conceptual-semantic relations**: **hypernym/hyponym**

also holonym/meronym

Also lexical relations, in particular lemmatization

Available at http://wordnet.princeton.edu
A WordNet example


Word to search for: bass

Display Options: (Select option to change) ➔ Change

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

Noun

- **S:** (n) bass (the lowest part of the musical range)
- **S:** (n) bass, bass part (the lowest part in polyphonic music)
- **S:** (n) bass, basso (an adult male singer with the lowest voice)
- **S:** (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
- **S:** (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S:** (n) bass, bass voice, basso (the lowest adult male singing voice)
- **S:** (n) bass (the member with the lowest range of a family of musical instruments)
- **S:** (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- **S:** (adj) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"
Hierarchical synset relations: nouns

**Hypernym/hyponym (between concepts)**
The more general ‘*meal*’ is a hypernym of the more specific ‘*breakfast*’

**Instance hypernym/hyponym (between concepts and instances)**
*Austen* is an instance hyponym of *author*

**Member holonym/meronym (groups and members)**
*professor* is a member meronym of (a university’s) *faculty*

**Part holonym/meronym (wholes and parts)**
*wheel* is a part meronym of (is a part of) *car*.

**Substance meronym/holonym (substances and components)**
*flour* is a substance meronym of (is made of) *bread*
Hierarchical synset relations: verbs

**Hypernym/troponym** (between events):
*travel/fly, walk/stroll*

*Flying* is a troponym of *traveling*:
it denotes a specific manner of *traveling*

**Entailment** (between events):
*snore/sleep*

*Snoring entails (presupposes)* *sleeping*
WordNet Hypernyms and Hyponyms

- **S: (n) bass** (the lowest part of the musical range)
  - *direct hypernym / inherited hypernym / sister term*
    - **S: (n) pitch** (the property of sound that varies with variation in the frequency of vibration)
    - **S: (n) sound property** (an attribute of sound)
    - **S: (n) property** (a basic or essential attribute shared by all members of a class) "a student"
    - **S: (n) attribute** (an abstraction belonging to or characteristic of an entity)
    - **S: (n) abstraction, abstract entity** (a general concept formed by extracting common properties)
    - **S: (n) entity** (that which is perceived or known or inferred to have it)
  - **S: (n) bass, bass part** (the lowest part in polyphonic music)
    - *direct hypernym / full hypernym*
      - **S: (n) ground bass** (a short melody in the bass that is constantly repeated)
      - **S: (n) figured bass, basso continuo, continuo, thorough bass** (a bass part written out in full and accented)
    - *direct hypernym / inherited hypernym / sister term*
      - **S: (n) part, voice** (the melody carried by a particular voice or instrument in polyphonic music) "he is the tenor of our times"
      - **S: (n) tune, melody, air, strain, melodic line, line, melodic phrase** (a succession of notes forming a musical idea)
      - **S: (n) music** (an artistic form of auditory communication incorporating instrumental or vocal elements)
      - **S: (n) auditory communication** (communication that relies on hearing)
      - **S: (n) communication** (something that is communicated by or to or between people)
      - **S: (n) abstraction, abstract entity** (a general concept formed by extracting common properties)
      - **S: (n) entity** (that which is perceived or known or inferred to have it)
WordNet-based Word Similarity
WordNet-based word similarity

There have been many attempts to exploit resources like WordNet to compute word (sense) similarities.

Classic approaches use the distance (path length) between synsets, possibly augmented with corpus statistics.

More recent (neural) approaches aim to learn (non-Euclidean) embeddings that capture the hierarchical structure of WordNet.
WordNet path lengths: examples and problems

Path length is just the distance between synsets

pathlen(nickel, dime) = 2  (nickel—coin—dime)
pathlen(nickel, money) = 5  (nickel—…—medium of exchange—money)
pathlen(nickel, budget) = 7  (nickel—…—medium of exchange—…—budget)

But do we really want the following?

pathlen(nickel, coin) < pathlen(nickel, dime)
pathlen(nickel, Richter scale) = pathlen(nickel, budget)
Problems with thesaurus-based similarity

We need to have a thesaurus! (not available for all languages)

We need to have a thesaurus that contains the words we’re interested in.

We need a thesaurus that captures a rich hierarchy of hypernyms and hyponyms.

Most thesaurus-based similarities depend on the specifics of the hierarchy that is implement in the thesaurus.
Learning hyponym relations

If we don’t have a thesaurus, can we learn that Corolla is a kind of car?

Certain phrases and patterns indicate hyponym relations:

- Hearst(1992) [Hearst patterns]
- Enumerations: cars such as the Corolla, the Civic, and the Vibe,
- Appositives: the Corolla, a popular car...

We can also learn these patterns if we have some seed examples of hyponym relations (e.g. from WordNet):

1. Take all hyponym/hypernym pairs from WordNet (e.g. car/vehicle)
2. Find all sentences that contain both, and identify patterns
3. Apply these patterns to new data to get new hyponym/hypernym pairs
Word Sense Disambiguation (WSD)
What does this word mean?

This **plant** needs to be **watered** each day.
 ⇒ **living plant**

This **plant** manufactures 1000 **widgets** each day.
 ⇒ **factory**

**Word Sense Disambiguation (WSD):**
Identify the sense of content words (nouns, verbs, adjectives) in context (assuming a fixed inventory of word senses).
Presumes the words to classify have a discrete set of senses.
Dictionary-based methods

We often don’t have a labeled corpus, but we might have a dictionary/thesaurus that contains glosses and examples:

**bank1**
Gloss: a financial institution that accepts deposits and channels the money into lending activities
Examples: “he cashed the check at the bank”, “that bank holds the mortgage on my home”

**bank2**
Gloss: sloping land (especially the slope beside a body of water)
Examples: “they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the current”
The Lesk algorithm

Simple, dictionary-based baseline for WSD
Basic idea: Compare the context with the dictionary definition of the sense.
   Assign the dictionary sense whose gloss and examples are most similar to the context in which the word occurs.

Compare the signature of a word in context with the signatures of its senses in the dictionary
Assign the sense that is most similar to the context

**Signature** = set of content words 
(in examples/gloss or in context)
**Similarity** = size of intersection of context signature and sense signature
Lesk algorithm

**bank1:**
Gloss: a financial institution that accepts deposits and channels the money into lending activities
Examples: “he cashed the check at the bank”, “that bank holds the mortgage on my home”

\[ \text{Signature}(\text{bank1}) = \{\text{financial, institution, accept, deposit, channel, money, lend, activity, cash, check, hold, mortgage, home}\} \]

**bank2:**
Gloss: sloping land (especially the slope beside a body of water)
Examples: “they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the current”

\[ \text{Signature}(\text{bank2}) = \{\text{slope, land, body, water, pull, canoe, sit, river, watch, current}\} \]

Target sentence: “The bank refused to give me a loan.”

Original signature: words in context \{refuse, give, loan\}
Augmented signature: add signatures of words in context (all senses) \{refuse, reject, request, ..., give, gift, donate, ..., loan, money, borrow, ...\}

Lesk algorithm: Pick the sense whose signature has greatest overlap to (augmented) signature of the target word
WSD as a learning problem

Supervised:
- You have a (large) corpus annotated with word senses
- Here, WSD is a standard supervised learning task:
  predict 1 of k senses for each occurrence of a word
  (depending on its context)

Semi-supervised (bootstrapping) approaches:
- You only have very little annotated data
  (and a lot of raw text)
- Here, WSD is a semi-supervised learning task
Implementing a WSD classifier

**Basic insight:** The sense of a word in a context depends on the words in its context.

**Features:**
- Which words in context: all words, all/some content words
- How large is the context? sentence, prev/following 5 words
- Do we represent context as **bag of words** (unordered set of words) or do we care about the **position** of words (preceding/following word)?
- Do we care about **POS tags**?
- Do we represent words as they occur in the text or as their **lemma** (dictionary form)?
Yarowsky’s weakly-supervised algorithm

1. **Initialization:**
   - Label a few seed examples *(that’s one form of supervision)*
   - Train an initial classifier on these seed examples

2. **Relabel:**
   - Label all unlabeled examples with the current classifier.
   - Add all examples that are labeled with high confidence to the labeled data set.
   - Apply **one-sense-per-discourse heuristic** to correct mistakes and get additional labeled examples *(that’s another form of supervision)*
     [Assume all occurrences of the same token (e.g. plant) in the same document have the same sense — this is true often enough that it can be very helpful, since it may be easy to label one occurrence correctly, and then you get the other labeled instances for free]

3. **Retrain:**
   - Train a new classifier on the new labeled data set.

4. **Repeat** 2. and 3. until convergence.

https://www.aclweb.org/anthology/P95-1026.pdf
Problems with word sense

Words can take on new meanings

**Metaphors**: bigger *fish* to fry

**Metonymy**: *The SUV* honked at me

[i.e. the SUV driver honked at me]

Word senses can be *modulated*

to identify different aspects of meaning:

- She oiled her bike [bike = bike chain]
- She dried off her bike. [bike = bike frame]
- Her bike goes like the wind [bike = the bike’s motion]

Kilgarriff: I don’t believe in Word Senses