

Lexical Semantics

Chapter 19

Lecture #13

November 2012

Largely based on Kathy Mckeown's lecture
which was based on Dan Jurafsky, Jim
Martin, and Chris Manning

1

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.**

2

Outline: Computational Lexical Semantics

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

3

Preliminaries

- What's a word?
 - Definitions we've used over the class: Types, tokens, stems, roots, uninflected 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
- Lemma – citation form – uninflected form (used to represent a lexeme). Need to do morphological parsing to get from wordform to lemma (lemmatization)
- Lemma is part-of-speech specific (e.g., table N and V)

4

Relationships between word meanings

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

5

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

6

Homonymy causes problems for NLP applications

- Text-to-Speech
 - Same orthographic form but different phonological form
 - Bass vs bass
 - Bow vs bow
 - Record vs record
- Information retrieval
 - Different meanings same orthographic form
 - QUERY: bat care
- Machine Translation
- Speech recognition

7

Polysemy

- The **bank** is constructed from red brick
I withdrew the money from the **bank**
- Are those the same sense?
- What about river **bank**?
- What about: The food **bank** is having a donation drive next week.
- Different senses but some more related than others...
- When two senses are related semantically we call it **polysemy** (rather than homonymy)

8

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)

9

Metaphor and Metonymy

- Specific types of polysemy
- Metaphor:
 - Germany will pull Slovenia out of its economic slump.
 - I spent 2 hours on that homework.
 - I put money into Google stock.
- Metonymy (use of one aspect of a concept or entity to refer to other aspects of the entity or to the entity itself)
 - The White House announced yesterday...
 - White House refers to the administration whose office is in the White House
 - This chapter talks about part-of-speech tagging
 - Bank (building) and bank (financial institution)

10

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?

11

Synonyms

- Words that have the same meaning in some or all contexts
 - Filbert / hazelnut
 - Couch / sofa
 - Big / large
 - Automobile / car
 - Vomit / throw up
 - Water / H₂O
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
 - If so they have the same **propositional meaning**

12

Synonyms

- But there are few (or no) examples of perfect synonym
 - 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 H₂O
 - Large coke versus *big coke

13

Some more terminology

- Lemmas and word forms
 - 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**
 - **Corpus** is the lemma for **corpora**
 - Specific surface forms carpets, sung, corpora are called **wordforms**
- The lemma bank has two senses:
 - Instead, a **bank** can hold the investments in...
 - 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

14

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

15

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 are at opposite ends of a scale (*long/short, fast/slow*)
 - Be reversives (describe a change of movement in opposite directions): *rise/fall, up/down*

16

Hyponym

- 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

17

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)

18

II. Wordnet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary

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

19

Wordnet

- Where it is:
- <http://wordnetweb.princeton.edu/perl/webwn>

20

Format of WordNet Entries

- The noun **bass** has 8 senses in wordnet:
- 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)
- **And 1 Adjective Sense:**
- 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"

21

WordNet Noun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> ¹ → <i>meal</i> ¹
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> ¹ → <i>lunch</i> ¹
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> ² → <i>professor</i> ¹
Has-Instance		From concepts to instances of the concept	<i>composer</i> ¹ → <i>Bach</i> ¹
Instance		From instances to their concepts	<i>Austen</i> ¹ → <i>author</i> ¹
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> ¹ → <i>crew</i> ¹
Part Meronym	Has-Part	From wholes to parts	<i>table</i> ² → <i>leg</i> ³
Part Holonym	Part-Of	From parts to wholes	<i>course</i> ¹ → <i>meal</i> ¹
Antonym		Opposites	<i>leader</i> ¹ → <i>follower</i> ¹

22

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> ³ → <i>travel</i> ⁵
Troponym	From a verb (event) to a specific manner elaboration of that verb	<i>walk</i> ¹ → <i>stroll</i> ¹
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> ¹ → <i>sleep</i> ¹
Antonym	Opposites	<i>increase</i> ¹ ↔ <i>decrease</i> ¹

23

24

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
```

25

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](#) (a person who is gullible and easy to take advantage of)
- Each of these senses share this same gloss
- Thus, for WordNet, the meaning of this sense of **chump** is this list.

26

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 the 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

27

Two variants of WSD task

- Lexical Sample task
 - Small pre-selected set of target words
 - And inventory of senses for each word
- 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

28

Approaches

- Supervised
- Semi-supervised
 - Unsupervised
 - Dictionary-based techniques
 - Selectional association
 - Lightly supervised
 - Bootstrapping
 - Preferred Selectional Association

29

Supervised Machine Learning Approaches

- Supervised machine learning approach:
 - A training corpus of ?
 - Used to train a classifier that can tag words in text
 - Just as in 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

30

Supervised WSD 1: WSD Tag

- What's a tag?

31

WordNet Bass

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

32

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	lubina	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...

33

Supervised WSD 2: Get a corpus

- Lexical sample task:
 - Line-hard-serve corpus -4000 examples of each
 - Interestcorpus -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

34

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?"

35

- Dishes
- Bass

36

- washing *dishes* .
- simple *dishes* including
- convenient *dishes* to
- of *dishes* and

- free *bass* with
- pound *bass* of
- and *bass* player
- his *bass* while

37

- “In our house, everybody has a career and none of them **includes washing dishes,**” he says.
- In her tiny kitchen at home, Ms. Chen works efficiently, stir-frying several simple *dishes*, **including braised pig’s ears** and chicken livers with green peppers.
- Post quick **and convenient dishes to fix** when your in a hurry.
- Japanese cuisine offers a great **variety of dishes and regional specialties**

38

- We need more good teachers –right now, there are only a half a dozen who can play **the free bass** with ease.
- Though still a far cry from the lake’s record **52-pound bass of a decade ago**, “you could fillet these fish again, and that made people very, very happy.” Mr. Paulson says.
- 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 again.
- Lowe **caught his bass while fishing** with pro Bill Lee of Killeen, Texas, who is currently in 144th place with two bass weighing 2-09.

39

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

How big should that window be?

40

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

41

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

42

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

43

Collocational

- Position-specific information about the words in the window
- guitar and bass player stand
- [guitar, NN, and, CC, player, NN, stand, VB]
- Wordn-2, POSn-2, wordn-1, POSn-1, Wordn+1 POSn+1...
- In other words, a vector consisting of
- [position n word, position n part-of-speech...]

44

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.

45

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...]

46

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...

47

WSD Evaluations and Baselines

- In vivo (end-to-end, extrinsic, task-based) versus in vitro (intrinsic as if a stand-alone system) evaluation
 - In vitro evaluation is most common now
 - Exact match accuracy
 - % of words tagged identically with manual sense tags
 - Usually evaluate using held-out data from same labeled corpus
 - Problems?
 - Why do we do it anyhow?
- Baselines
 - Most frequent sense
 - The Lesk algorithm (choose the sense whose dictionary gloss or definition shares the most words with the target word's neighborhood.)

48

Most Frequent Sense

- WordNet senses are order in frequency order
- So “most frequent sense” in WordNet = “take the first sense”

Freq	Synset	Gloss
338	plant ¹ , works, industrial plant	buildings for carrying on industrial labor
207	plant ² , flora, plant life	a living organism lacking the power of locomotion
2	plant ³	something planted secretly for discovery by another
0	plant ⁴	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

49

Ceiling

- Human-inter-annotator agreement
 - Compare annotations of two humans
 - On same data
 - Given same tagging guidelines
- Human agreements on all-words corpora with WordNet style senses
 - 75%-80%

50