CS114 Lecture 16

Lexical Semantics Continued

March 24, 2014

Professor Meteer

Thanks for Jurafsky & Martin & Prof. Pustejovksy for slides

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

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**

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

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 that can discriminate between the various senses
- 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…

Classifiers

- The choice of technique, in part, depends on the set of features that have been used
  - Some techniques work better/worse with features with numerical values
  - Some techniques work better/worse with features that have large numbers of possible values
    - For example, the feature **the word to the left** has a fairly large number of possible values

Naïve Bayes

- The sense with the highest probability given the feature vector

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \; P(s \mid \vec{f})$$

- Rewrite with Bayes

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \; \frac{P(\vec{f} \mid s)\, p(s)}{p(\vec{f})}$$

- Remove denominator

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \; P(\vec{f} \mid s) P(s)$$

- Assume independence of the features

$$P(\vec{f} \mid s) \approx \prod_{j=1}^{n} P(f_j \mid s)$$

- Final:

$$\hat{s} \approx \underset{s \in S}{\operatorname{argmax}} \; P(s) \prod_{j=1}^{n} P(f_j \mid s)$$

Naïve Bayes

- $P(s)$ … just the prior probability of that sense.
  - Just as with part of speech tagging, not all senses will occur with equal frequency
  - $P(si) = count(si,wj)/count(wj)$
- $P(fj|s)$… conditional probability of some particular feature/value combination given a particular sense
  - $P(fj|s) = count(fj,s)/count(s)$
- You can get both of these from a tagged corpus with the features encoded

Naïve Bayes Test

- On a corpus of examples of uses of the word line, naïve Bayes achieved about 73% correct

- Good?

Decision Lists: another popular method

- A case statement....

Rule		Sense
fish within window	$\Rightarrow$	**bass**^{1}
striped bass	$\Rightarrow$	**bass**^{1}
guitar within window	$\Rightarrow$	**bass**^{2}
bass player	$\Rightarrow$	**bass**^{2}
piano within window	$\Rightarrow$	**bass**^{2}
tenor within window	$\Rightarrow$	**bass**^{2}
sea bass	$\Rightarrow$	**bass**^{1}
play/V bass	$\Rightarrow$	**bass**^{2}
river within window	$\Rightarrow$	**bass**^{1}
violin within window	$\Rightarrow$	**bass**^{2}
salmon within window	$\Rightarrow$	**bass**^{1}
on bass	$\Rightarrow$	**bass**^{2}
bass are	$\Rightarrow$	**bass**^{1}

Learning Decision Lists

- Restrict the lists to rules that test a single feature (1-decisionlist rules)
- Evaluate each possible test and rank them based on how well they work.
- Glue the top-N tests together and call that your decision list.

Yarowsky

- On a binary (homonymy) distinction used the following metric to rank the tests

$$\frac{P(\text{Sense}_1 \mid Feature)}{P(\text{Sense}_2 \mid Feature)}$$

- Ratio tells us how discriminating this feature is
- Order the tests by the log-likelihood ratio
- This gives about 95% on this test…

WSD Evaluations and baselines

- *In vivo* versus *in vitro* 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

Most Frequent Sense

- Wordnet senses are ordered in frequency order
- So “most frequent sense” in wordnet = “take the first sense”
- Sense frequencies come from SemCor

Freq	Synset	Gloss
338	plant1, works, industrial plant	buildings for carrying on industrial labor
207	plant-, flora, plant life	a living organism lacking the power of locomotion
2	plant	something planted secretly for discovery by another
0	plant	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

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%

WSD: Dictionary/Thesaurus methods

- The Lesk Algorithm
- Selectional Restrictions and Selectional Preferences

Simplified Lesk

The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

Given the following two WordNet senses:

$Bank_1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$Bank_2$	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 currents"

Simplified Lesk

- Count the overlap between the context and the dictionary definition
  - Sentence: "The bank can guarantee deposits will eventually cover future tuition costs because it invest in adjustable-rate mortgage securities

$Bank_1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$Bank_2$	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 currents"

Original Lesk: pine cone

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

Corpus Lesk

- Add corpus examples to glosses and examples
- The best performing variant

Bootstrapping

- What if you don't have enough data to train a system…
- Bootstrap
  - Pick a word that you as an analyst think will co-occur with your target word in particular sense
  - Grep through your corpus for your target word and the hypothesized word
  - Assume that the target tag is the right one

Bootstrapping

- For bass
  - Assume play occurs with the music sense and fish occurs with the fish sense

Sentences extracting using “fish” and “play”

We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease. An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps. When the New Jersey Jazz Society, in a fund-raiser for the American Jazz Hall of Fame, honors this historic night next Saturday, Harry Goodman, Mr. Goodman's brother and **bass play**er at the original concert, will be in the audience with other family members.
The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper. And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny. 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.

Yarowsky Bootstrapping

- Label a small set of examples and train a “decision list” classifier on these examples
  - For plant, “life” is Sense-A and “manufacturing” is Sense-B
- Apply the classifier to all of the instances.
  - Select those with a high score and add them to the training set
- Create a new decision set classifier
  - For Sense-A: life, animal, microscopic
  - For Sense-B: employee, equipment
- Repeat until all the instances are labeled or performance doesn’t improve

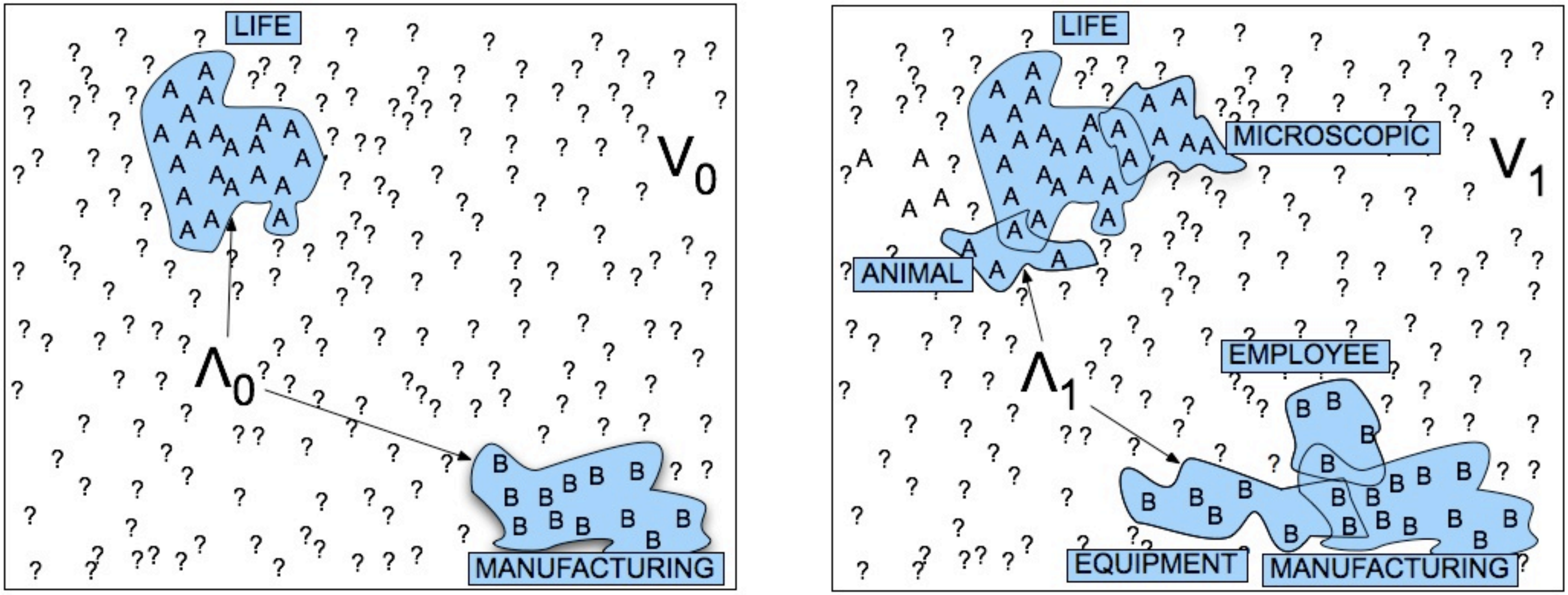
Where do the seeds come from?

1) Hand labeling
2) “One sense per discourse”:
   - The sense of a word is highly consistent within a document - Yarowsky (1995)
   - True for topic dependent words
   - Not so true for other POS like adjectives and verbs, e.g. make, take
   - Krovetz (1998) “More than one sense per discourse” argues it isn’t true at all once you move to fine-grained senses
3) One sense per collocation:
   - A word reoccurring in collocation with the same word will almost surely have the same sense.

Slide adapted from Chris Manning

Stages in the Yarowsky bootstrapping algorithm

Plant



(a) (b)

Problems

- Given these general ML approaches, how many classifiers do I need to perform WSD robustly
  - One for each ambiguous word in the language
- How do you decide what set of tags/labels/senses to use for a given word?
  - Depends on the application

WordNet Bass

- Tagging with this set of senses is an impossibly hard task that's probably overkill for any realistic application

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)

Senseval History

- ACL-SIGLEX workshop (1997)
  - Yarowsky and Resnik paper
- SENSEVAL-I (1998)
  - Lexical Sample for English, French, and Italian
- SENSEVAL-II (Toulouse, 2001)
  - Lexical Sample and All Words
  - Organization: Kilkgarriff (Brighton)
- SENSEVAL-III (2004)
- SENSEVAL-IV -> SEMEVAL (2007)

WSD Performance

- Varies widely depending on how difficult the disambiguation task is
- Accuracies of over 90% are commonly reported on some of the classic, often fairly easy, WSD tasks (pike, star, interest)
- Senseval brought careful evaluation of difficult WSD (many senses, different POS)
- Senseval 1: more fine grained senses, wider range of types:
  - Overall: about 75% accuracy
  - Nouns: about 80% accuracy
  - Verbs: about 70% accuracy