Word Relations and Word Sense Disambiguation

Slides adapted from Dan Jurafsky, Jim Martin and Chris Manning

Homework Questions?
Schedule

- This week
  - Finish semantics
  - Begin machine learning for NLP
  - Review for midterm

- Midterm
  - October 27th,
  - Where: 1024 Mudd (here)
  - When: Class time, 2:40-4:00
  - Will cover everything through semantics
  - A sample midterm will be posted
  - Includes multiple choice, short answer, problem solving

- October 29th
  - Bob Coyne and Words Eye: Not to be missed!

- TBD: Class outing to *Where the Wild Things Are*

Recap on WSD

- A subset of WordNet sense representation commonly used
- WordNet provides many relations that capture meaning
- To do WSD, need a training corpus tagged with senses
- Naïve Bayes approach to learning the correct sense
  - Probability of a specific sense given a set of features
  - Collocational features
  - Bag of words

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

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

- 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	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

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%

Unsupervised Methods WSD: Dictionary/Thesaurus methods

- The Lesk Algorithm
- Selectional Restrictions

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"

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

Disambiguation via Selectional Restrictions

- "Verbs are known by the company they keep"
  - Different verbs select for different thematic roles

    wash the *dishes* (takes washable-thing as patient)

    serve delicious *dishes* (takes food-type as patient)
- Method: another semantic attachment in grammar
  - Semantic attachment rules are applied as sentences are syntactically parsed, e.g.

    VP --> V NP

    V→ serve <theme> {theme:food-type}
  - Selectional restriction violation: no parse

- But this means we must:
  - Write selectional restrictions for each sense of each predicate – or use FrameNet
    - Serve alone has 15 verb senses
  - Obtain hierarchical type information about each argument (using WordNet)
    - How many hypernyms does dish have?
    - How many words are hyponyms of dish?
- But also:
  - Sometimes selectional restrictions don't restrict enough (Which dishes do you like?)
  - Sometimes they restrict too much (Eat dirt, worm! I'll eat my hat!)
- Can we take a statistical approach?

Semi-supervised 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.

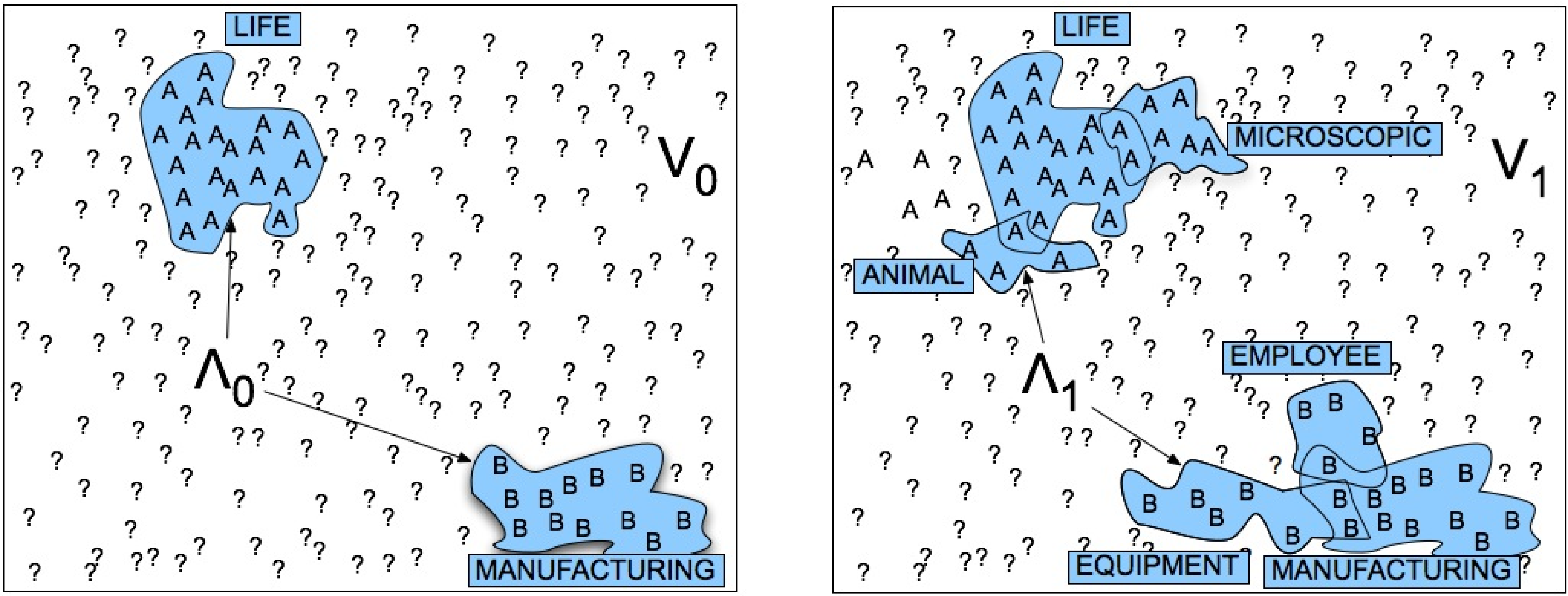
And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

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.

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.

Stages in the Yarowsky bootstrapping algorithm



(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

Summary

- Lexical Semantics
  - Homonymy, Polysemy, Synonymy
  - Thematic roles
- Computational resource for lexical semantics
  - WordNet
- Task
  - Word sense disambiguation

Statistical NLP

- Machine Learning for NL Tasks
  - Some form of classification

- Experiment with the impact of different kinds of NLP knowledge

What useful things can we do with this knowledge?

- Find sentence boundaries, abbreviations
- Sense disambiguation
  Find Named Entities (person names, company names, telephone numbers, addresses,…)
- Find topic boundaries and classify articles into topics
- Identify a document’s author and their opinion on the topic, pro or con
- Answer simple questions (factoids)
- Do simple summarization

Corpus

- Find or annotate a corpus
- Divide into training and test

Next, we pose a question…the dependent variable

- Binary questions:
  - Is this word followed by a sentence boundary or not?
  - A topic boundary?
  - Does this word begin a person name? End one?
  - Should this word or sentence be included in a summary?
- Classification:
  - Is this document about medical issues? Politics? Religion? Sports? …
- Predicting continuous variables:
  - How loud or high should this utterance be produced?

Finding a suitable corpus and preparing it for analysis

- Which corpora can answer my question?
  - Do I need to get them labeled to do so?
- Dividing the corpus into training and test corpora
  - To develop a model, we need a training corpus
    - overly narrow corpus: doesn’t generalize
    - overly general corpus: don't reflect task or domain
  - To demonstrate how general our model is, we need a test corpus to evaluate the model
    - Development test set vs. held out test set
  - To evaluate our model we must choose an evaluation metric
    - Accuracy
    - Precision, recall, F-measure,…
    - Cross validation

Then we build the model…

- Identify the dependent variable: what do we want to predict or classify?
  - Does this word begin a person name? Is this word within a person name?
  - Is this document about sports? stocks? Health? International news? ???
- Identify the independent variables: what features might help to *predict* the dependent variable?
  - What words are used in the document?
  - Does ‘hockey’ appear in this document?
  - What is this word’s POS? What is the POS of the word before it? After it?
  - Is this word capitalized? Is it followed by a ‘.’?
  - Do terms play a role? (e.g., “myocardial infarction”, “stock market,” “live stock”)
  - How far is this word from the beginning of its sentence?
- Extract the values of each variable from the corpus by some automatic means

A Sample Feature Vector for Sentence-Ending Detection

WordID	POS	Cap?	, After?	Dist/Sbeg	End?
Clinton	N	y	n	1	n
won	V	n	n	2	n
easily	Adv	n	y	3	n
but	Conj	n	n	4	n

An Example: Genre Identification

- Automatically determine
  - Short story
    - Aesop's Fable
    - Fairy Tale
    - Children's story
  - Poetry
  - News
  - Email

Corpus?

- British National Corpus
  - Poetry
  - Fiction
  - Academic Prose
  - Non-academic Prose
- http://aesopfables.com
- Enron corpus: http://www.cs.cmu.edu/~enron/

Features?

The Ant and the Dove

AN ANT went to the bank of a river to quench its thirst, and being carried away by the rush of the stream, was on the point of drowning. A Dove sitting on a tree overhanging the water plucked a leaf and let it fall into the stream close to her. The Ant climbed onto it and floated in safety to the bank. Shortly afterwards a birdcatcher came and stood under the tree, and laid his lime-twigs for the Dove, which sat in the branches. The Ant, perceiving his design, stung him in the foot. In pain the birdcatcher threw down the twigs, and the noise made the Dove take wing.

One good turn deserves another

First Fig

My candle burns at both ends;
  It will not last the night;
But ah, my foes, and oh, my friends--
  It gives a lovely light!

Edna St. Vincent Millay

CHICAGO SUN-TIMES
suntimes.com Member of the Sun-Times News Group

BARACK OBAMA

Obama makes surprise appearance

HISTORIC DAY | Nominee makes surprise appearance to hail Biden, salute Clintons

August 28, 2008

BY ABDON M. PALLASCH AND DAVE MCKINNEY Staff Reporters

DENVER -- Barack Obama walked onto the stage of the Democratic National Convention late Wednesday, just hours after he walked into the history books.

The skinny guy from the South Side of Chicago became the first African American to head a major party presidential ticket when delegates unanimously nominated him by a voice vote in late afternoon.

RELATED STORIES
• It's time he also reaches out to urban blacks
• FEUDING DEMS HUGGING IT OUT
• Obama lacks experience? Bill says not anymore
• The hugfest . . .
• Hartmarx suits Obama

Then minutes before the convention closed for the night, Obama made a surprise appearance in the convention hall to congratulate running mate Joe Biden on his fiery speech accepting the vice presidential nomination.

"Hello, Democrats. I just wanted to come out here with a little something to say," Obama told the cheering crowd. "I want everybody to now understand why I am so proud to have Joe Biden ... and the whole Biden family with me on this journey to

SUN-TIMES BLOGS ::

BACKTALK BLOG
• Palin's Pastor Problems (9/12/2008 16:44:46 PM)

CONQUERING CANCER AND HEART FAILURE BLOG
• Hallelujah! I'm back to work. (9/20/2008 03:51:20 AM)

CONVENTIONAL WISDOM BLOG
• Audio slideshow: McCain's acceptance speech

Email

- Dear Professor, I'll see you at 6 pm then.
  Regards, Madhav

- On Wed, Sep 24, 2008 at 12:06 PM, Kathy McKeown
  <kathy@cs.columbia.edu> wrote:
  > I am on the eexamining committee of a candidacy exam from 4-5. That is the
  > reason I changed my office hours. If you come right at 6, should be OK. It
  > is important that you stop by.
  > > Kathy

  > > Madhav Krishna wrote:
  >> >> Dear Professor,
  >> >> Can I come to your office between, say, 4-5 pm today? Google has a
  >> >> tech talk on campus today starting at 5 pm -- I would like to attend.
  >> >> Regards.

Genre Identification Approaches

- *Kessler, Nunberg, and Schutze, Automatic Detection of Text Genre, EACL 1997, Madrid, Spain.*

- Karlgren and Cutting, Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of Coling 94,* Kyoto, Japan.

Why Genre Identification?

- Parsing accuracy can be increased
  - E.g., recipes
- POS tagging accuracy can be increased
  - E.g., “trend” as a verb
- Word sense disambiguation
  - E.g., “pretty” in informal genres
- Information retrieval
  - Allow users to more easily sort through results

What is genre?

- Is genre a single property or a multi-dimensional space of properties?
- Class of text
  - Common function
  - Function characterized by formal features
  - Class is extensible
    - Editorial vs. persuasive text
- Genre facets
  - BROW
    - Popular, middle, upper-middle, high
  - NARRATIVE
    - Yes, no
  - GENRE
    - Reportage, editorial, scitech, legal, non-fiction, fiction

Corpus

- 499 texts from the Brown corpus
  - Randomly selected
- Training: 402 texts
- Test: 97 texts
  - Selected so that equal representation of each facet

Features

- Structural Cues
  - Passives, nominalizations, topicalized sentences, frequency of POS tags
  - Used in Karlgren and Cutting
- Lexical Cues
  - Mr., Mrs. (in papers like the NY Times)
  - Latinate affixes (should signify high brow as in scientific papers)
  - Dates (appear frequently in certain news articles)
- Character Cues
  - Punctuation, separators, delimiters, acronyms
- Derivative Cues
  - Ratios and variation metrics derived from lexical, character and structural cues
  - Words per sentence, average word length, words per token
  - 55 in total used

- *Kessler et al hypothesis: The surface cues will work as well as the structural cues*

Machine Learning Techniques

- Logistic Regression
- Neural Networks
  - To avoid overfitting given large number of variables
  - Simple perceptron
  - Multi-layer perceptron

Baselines

- Karlgren and Cutting
  - Can they do better or, at least, equivalent, using features that are simpler to compute?
- Simple baseline
  - *Choose the majority class*
  - Another possibility: random guess among the k categories
    - 50% for narrative (yes,no)
    - 1/6 for genre
    - ¼ for brow

Table 1: Classification Results for All Facets.

Facet	Baseline	LR (Surf.)		2LP		3LP		LR (Struct.)	
		All	Sel.	All	Sel.	All	Sel.	All	Sel.
Narrative	54	78	80	82	82	86	82	78	80
Genre	33	61	66	75	79	71	74	66	62
Brow	32	44	46	47	—	54	—	46	53

Note. Numbers are the percentage of the evaluation subcorpus ($N = 97$) which were correctly assigned to the appropriate facet level: the Baseline column tells what percentage would be correct if the machine always guessed the most frequent level. LR is Logistic Regression, over our surface cues (Surf.) or Karlgren and Cutting's structural cues (Struct.): 2LP and 3LP are 2- or 3-layer perceptrons using our surface cues. Under each experiment, All tells the results when all cues are used, and Sel. tells the results when for each level one selects the most discriminating cues. A dash indicates that an experiment was not run.

Table 2: Classification Results for Each Facet Level.

Levels	Baseline	LR (Surf.) All	LR (Surf.) Sel.	2LP All	3LP All	LR (Struct.) All	LR (Struct.) Sel.
Genre							
Rep	81	89*	88	94*	94*	90*	90*
Edit	81	75	75	74	80	79	77
Legal	95	96	96	95	95	93	93
Scitech	94	100*	96	99*	94	93	96
Nonfict	67	67	68	78*	67	73	74
Fict	81	93*	96*	99*	81	96*	96*
Brow							
Popular	74	74	75	74	74	72	73
Middle	68	66	67	64	64	58	64
Uppermiddle	88	74	78	86	88	79	82
High	70	84*	88*	89*	90*	85*	86*

Note. Numbers are the percentage of the evaluation subcorpus ($N = 97$) which was correctly classified on a binary discrimination task. The Baseline column tells what percentage would be got correct by guessing No for each level. Headers have the same meaning as in Table 1.

* means significantly better than Baseline at $p < .05$, using a binomial distribution (N=97, p as per first column).

Confusion Matrix

Table 3: Genre Binary Level Classification Results by Genre Level.

Actual	Guess						
	Rep	Edit	Legal	Scitech	Nonfict	Fict	N
Rep	83	6	0	0	11	0	18
Edit	17	61	0	0	17	6	18
Legal	20	0	20	0	60	0	5
Scitech	0	0	0	83	17	0	6
Nonfict	3	34	0	6	47	9	32
Fict	0	6	0	0	0	94	18

Note. Numbers are the percentage of the texts actually belonging to the GENRE level indicated in the first column that were classified as belonging to each of the GENRE levels indicated in the column headers. Thus the diagonals are correct guesses, and each row would sum to 100%, but for rounding error.

Discussion

- All of the facet classifications significantly better than baseline
- Component analysis
  - Some genres better than other
    - Significantly better on reportage and fiction
    - Better, but not significantly so on non-fiction and scitech
      - Infrequent categories in the Brown corpus
    - Less well for editorial and legal
      - Genres that are hard to distinguish
- Good performance on brow stems from ability to classify in the high brow category
- Only a small difference between structural and surface cues