# CSCI 5832
# Natural Language Processing

Jim Martin
Lecture 21



## Today 4/8

- Finish WSD
- Start on IE (Chapter 22)



## WSD and Selection Restrictions

- Ambiguous arguments
  - Prepare a dish
  - Wash a dish
- Ambiguous predicates
  - Serve Denver
  - Serve breakfast
- Both
  - Serves vegetarian dishes

## WSD and Selection Restrictions

- This approach is complementary to the compositional analysis approach.
  - You need a parse tree and some form of predicate-argument analysis derived from
    - The tree and its attachments
    - All the word senses coming up from the lexemes at the leaves of the tree
    - Ill-formed analyses are eliminated by noting any selection restriction violations



## Problems

- As we saw last time, selection restrictions are violated all the time.
- This doesn't mean that the sentences are ill-formed or preferred less than others.
- This approach needs some way of categorizing and dealing with the various ways that restrictions can be violated



## Supervised ML Approaches

- That's too hard… try something empirical
- In supervised machine learning approaches, a training corpus of words tagged in context with their sense is used to train a classifier that can tag words in new text (that reflects the training text)

## 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)



## Representations

- Most supervised ML approaches require a very simple representation for the input training data.
  - Vectors of sets of feature/value pairs
    - I.e. files of comma-separated values
- So our first task is to extract training data from a corpus with respect to a particular instance of a target word
  - This typically consists of a characterization of the window of text surrounding the target

## Representations

- This is where ML and NLP intersect
  - If you stick to trivial surface features that are easy to extract from a text, then most of the work is in the ML system
  - If you decide to use features that require more analysis (say parse trees) then the ML part may be doing less work (relatively) if these features are truly informative



## Surface Representations

- Collocational and co-occurrence information
  - Collocational
    - Encode features about the words that appear in specific positions to the right and left of the target word
      - Often limited to the words themselves as well as they're part of speech
  - Co-occurrence
    - Features characterizing the 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, CJC, player, NN, stand, VVB]
  - In other words, a vector consisting of
  - [position n word, position n part-of-speech…]



## Co-occurrence

- 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]



## Classifiers

- Once we cast the WSD problem as a classification problem, then all sorts of techniques are possible
  - Naïve Bayes (the right thing to try first)
  - Decision lists
  - Decision trees
  - MaxEnt
  - 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

- Argmax P(sense|feature vector)
- Rewriting with Bayes and assuming independence of the features



## Naïve Bayes

- P(s) … just the prior of that sense.
  - Just as with part of speech tagging, not all senses will occur with equal frequency
- $P(v_j|s)$… conditional probability of some particular feature/value combination given a particular sense
- 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?

## 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)



## Semantic Analysis

- When we covered semantic analysis in Chapter 18, we focused on
  - The analysis of single sentences
  - A deep approach that could, in principle, be used to extract considerable information from each sentence
    - Predicate-argument structure
    - Quantifier scope
    - Etc.
  - And a tight coupling with syntactic analysis

## Semantic Analysis

- Unfortunately, when released in the wild such approaches have difficulties with
  - Speed... Deep syntactic and semantic analysis of each sentence is too slow for many applications
    - Transaction processing where large amounts of newly encountered text has to be analysed
      - Blog analysis
      - Question answering
      - Summarization
  - Coverage... Real world texts tend to strain both the syntactic and semantic capabilities of most systems



## Information Extraction

- So just as we did with partial/parsing and chunking for syntax, we can look for more lightweight techniques that get us most of what we might want in a more robust manner.
  - Figure out the entities (the players, props, instruments, locations, etc. in a text)
  - Figure out how they're related
  - Figure out what they're all up to
  - And do each of those tasks in a loosely-coupled data-driven manner



## Information Extraction

- Ordinary newswire text is often used in typical examples.
  - And there's an argument that there are useful applications there
- The real interest/money is in specialized domains
  - Bioinformatics
  - Patent analysis
  - Specific market segments for stock analysis
  - Intelligence analysis
  - Etc.

## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York



## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.



## Named Entity Recognition

- Find the named entities and classify them by type.
- Typical approach
  - Acquire training data
  - Encode using IOB labeling
  - Train a sequential supervised classifier
  - Augment with pre- and post-processing using available list resources (census data, gazeteers, etc.)

## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York



## Relation Extraction

- Basic task: find all the classifiable relations among the named entities in a text (populate a database)…
  - Employs
    - { <American, Tim Wagner> }
  - Part-Of
    - { <United, UAL>, {American, AMR} >



## Relation Extraction

- Typical approach:
  For all pairs of entities in a text
  - Extract features from the text span that just covers both of the entities
    - Use a binary classifier to decide if there is likely to be a relation
    - If yes: then apply each of the known classifiers to the pair to decide which one it is
- Use supervised ML to train the required classifiers from an annotated corpus

## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York



## Event Detection

- Find and classify all the events in a text.
  - Most verbs introduce events/states
    - But not all (*give* a kiss)
  - Nominalizations often introduce events
    - *Collision, destruction, the running...*



## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York

## Temporal and Numerical Expressions

- Temporals
  - Find all the temporal expressions
  - Normalize them based on some reference point
- Numerical Expressions
  - Find all the expressions
  - Classify by type
  - Normalize



## Information Extraction

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York



## Template Analysis

- Many news stories have a script-like flavor to them. They have fixed sets of expected events, entities, relations, etc.
- Template, schemas or script processing involves:
  - Recognizing that a story matches a known script
  - Extracting the parts of that script

## Information Extraction Summary

- Named entity recognition and classification
- Coreference analysis
- Temporal and numerical expression analysis
- Event detection and classification
- Relation extraction
- Template analysis



## Next Time

- Rest of Chapter 22
  - More details
    - NER, relations, and templates
  - Bioinformatic examples