NLP

Introduction to NLP

Word Sense Disambiguation

Introduction

- Polysemy
  - Words have multiple senses
- Example
  - Let’s have a drink in the bar
  - I have to study for the bar
  - Bring me a chocolate bar
- Homonymy
  - May I come in?
  - Let’s meet again in May
- Part of speech ambiguity
  - Joe won the first round
  - Joe has a round toy

Senses Of The Word "Bar"

- S: (n) barroom, **bar**, saloon, ginmill, taproom (a room or establishment where alcoholic drinks are served over a counter) *"he drowned his sorrows in whiskey at the bar"*
- S: (n) **bar** (a counter where you can obtain food or drink) *"he bought a hot dog and a coke at the bar"*
- S: (n) **bar** (a rigid piece of metal or wood; usually used as a fastening or obstruction or weapon) *"there were bars in the windows to prevent escape"*
- S: (n) measure, **bar** (musical notation for a repeating pattern of musical beats) *"the orchestra omitted the last twelve bars of the song"*
- S: (n) **bar** (an obstruction (usually metal) placed at the top of a goal) *"it was an excellent kick but the ball hit the bar"*
- S: (n) prevention, **bar** (the act of preventing) *"there was no bar against leaving"; "money was allocated to study the cause and prevention of influenza"*
- S: (n) **bar** ((meteorology) a unit of pressure equal to a million dynes per square centimeter) *"unfortunately some writers have used bar for one dyne per square centimeter"*
- S: (n) **bar** (a submerged (or partly submerged) ridge in a river or along a shore) *"the boat ran aground on a submerged bar in the river"*
- S: (n) legal profession, **bar**, legal community (the body of individuals qualified to practice law in a particular jurisdiction) *"he was admitted to the bar in New Jersey"*
- S: (n) stripe, streak, **bar** (a narrow marking of a different color or texture from the background) *"a green toad with small black stripes or bars"; "may the Stars and Stripes forever wave"*
- S: (n) cake, **bar** (a block of solid substance (such as soap or wax)) *"a bar of chocolate"*
- S: (n) Browning automatic rifle, **BAR** (a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War)
- S: (n) **bar** (a horizontal rod that serves as a support for gymnasts as they perform exercises)
- S: (n) **bar** (a heating element in an electric fire) *"an electric fire with three bars"*
- S: (n) **bar** ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried) *"spectators were not allowed past the bar"*

Word Sense Disambiguation

- Task
  - given a word
  - and its context
  - determine which sense it is
- Use for Machine Translation
  - e.g., translate “play” into Spanish
  - play the violin = tocar el violín
  - play tennis = jugar al tenis
- Other uses
  - Accent restoration (cote)
  - Text to speech generation (lead)
  - Spelling correction (aid/aide)
  - Capitalization restoration (Turkey)

Dictionary Method (Lesk)

- Match sentences to dictionary definitions
- Examples of plant (m-w.com):
  - plant_1 = a living thing that grows in the ground, usually has leaves or flowers, and needs sun and water to survive
  - plant_2 = a building or factory where something is made
- Examples of leaf
  - leaf_1 = a lateral outgrowth from a plant stem that is typically a flattened expanded variably shaped greenish organ, constitutes a unit of the foliage, and functions primarily in food manufacture by photosynthesis
  - leaf_2 = a part of a book or folded sheet containing a page on each side
- Find the pair of meanings that have the most overlapping definitions
  - "The *leaf* is the food making factory of green *plants.*"

Decision Lists (Yarowsky)

- Method introduced by Yarowsky (1994)
- Two senses per word
- Ordered rules: collocation -> sense
- Formula

Decision Lists (Yarowsky)

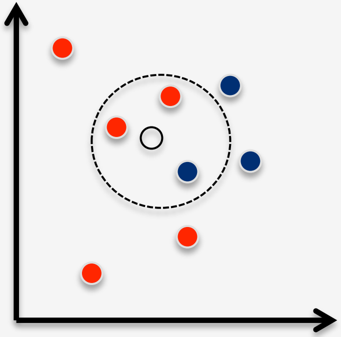
- *fish* within window -> bass1
- *striped bass* -> bass1
- *guitar* within window -> bass2
- *bass player -> bass2*
- Play/V bass -> *bass2*

Classification Features

- Adjacent words (collocations)
  - e.g., chocolate bar, bar exam, bar stool, bar fight, foreign aid, presidential aide
- Position
  - e.g., plant pesticide vs. pesticide plant
- Adjacent parts of speech
- Nearby words
  - e.g., within 10 words
- Syntactic information
  - e.g., object of the verb “play”
- Topic of the text

Classification Methods

- K-nearest neighbor (memory-based)
- Using Euclidean distance
- Find the k most similar examples and return the majority class for them



Bootstrapping

- Start with two senses and seeds for each sense
  - e.g., plant1:leaf, plant2:factory
- Use these seeds to label the data using a supervised classifier (decision list)
- Add some of the newly labeled examples to the training data
- Repeat until no more examples can be labeled

Bootstrapping

- Two principles:
  - one sense per collocation
  - one sense per discourse (e.g., document)

Training Data for WSD

- Senseval/Semcor
  - http://www.senseval.org/senseval3
  - Lexical Sample
  - All words
  - Available for many languages
- Pseudo-words
  - E.g., banana/door
- Multilingual corpora
  - Aligned at the sentence level
  - Use the translations as an indication of sense

Senseval-1 Evaluation

- Metric
  - A = number of assigned senses
  - C = number of words assigned correct senses
  - T = total number of test words
  - Precision = C/A; Recall = C/T
- Results
  - best recall around 77P/77R
  - human lexicographer 97P/96R
  - most common sense 57P/50R (decent but depends on domain)

NLP