# ISE 540 Text Analytics

Mayank Kejriwal

**Advanced Topic:** Word Sense Disambiguation (WSD)

# Lexical Ambiguity

- Most words in natural languages have multiple possible meanings.
  - “pen” (noun)
    - The dog is in the pen.
    - The ink is in the pen.
  - “take” (verb)
    - Take one pill every morning.
    - Take the first right past the stoplight.
- Syntax helps distinguish meanings for different parts of speech of an ambiguous word.
  - “conduct” (noun or verb)
    - John’s conduct in class is unacceptable.
    - John will conduct the orchestra on Thursday.

# Motivation for Word Sense Disambiguation (WSD)

- Many tasks in natural language processing require disambiguation of ambiguous words.
  - Question Answering
  - Information Retrieval
  - Machine Translation
  - Text Mining
  - Phone Help Systems
- Understanding how people disambiguate words is an interesting problem that can provide insight in psycholinguistics.

# Sense Inventory

- What is a “sense” of a word?
  - Homonyms (disconnected meanings)
    - bank: financial institution
    - bank: sloping land next to a river
  - Polysemes (related meanings with joint etymology)
    - bank: financial institution as corporation
    - bank: a building housing such an institution
- Sources of sense inventories
  - Dictionaries
  - Lexical databases

# WordNet

- A detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144,000 English words.
- Nouns, adjectives, verbs, and adverbs grouped into about 109,000 synonym sets called *synsets*.

# WordNet Synset Relationships

- Antonym: front → back
- Attribute: benevolence → good (noun to adjective)
- Pertainym: alphabetical → alphabet (adjective to noun)
- Similar: unquestioning → absolute
- Cause: kill → die
- Entailment: breathe → inhale
- Holonym: chapter → text (part to whole)
- Meronym: computer → cpu (whole to part)
- Hyponym: plant → tree (specialization)
- Hypernym: apple → fruit (generalization)

# EuroWordNet

- WordNets for
  - Dutch
  - Italian
  - Spanish
  - German
  - French
  - Czech
  - Estonian

# WordNet Senses

- WordNets senses (like many dictionary senses) tend to be very fine-grained.
- “play” as a verb has 35 senses, including
  - play a role or part: “Gielgud played Hamlet”
  - pretend to have certain qualities or state of mind: “John played dead.”
- Difficult to disambiguate to this level for people and computers. Only expert lexicographers are perhaps able to reliably differentiate senses.
- Not clear such fine-grained senses are useful for NLP.
- Several proposals for grouping senses into coarser, easier to identify senses (e.g. homonyms only).

# Senses Based on Needs of Translation

- Only distinguish senses that are translate to different words in some other language.
  - play: tocar vs. jugar
  - know: conocer vs. saber
  - be: ser vs. estar
  - leave: salir vs dejar
  - take: llevar vs. tomar vs. sacar
- May still require overly fine-grained senses
  - river in French is either:
    - fleuve: flows into the ocean
    - rivière: does not flow into the ocean

# Learning for WSD

- Assume part-of-speech (POS), e.g. noun, verb, adjective, for the target word is determined.
- Treat as a classification problem with the appropriate potential senses for the target word given its POS as the categories.
- Encode context using a set of features to be used for disambiguation.
- Train a classifier on labeled data encoded using these features.
- Use the trained classifier to disambiguate future instances of the target word given their contextual features.

# Feature Engineering

- The success of machine learning requires instances to be represented using an effective set of features that are correlated with the categories of interest.
- Feature engineering can be a laborious process that requires substantial human expertise and knowledge of the domain.
- In NLP it is common to extract many (even thousands of) potentially features and use a learning algorithm that works well with many relevant and irrelevant features.

# Issues in WSD

- What is the right granularity of a sense inventory?
- Integrating WSD with other NLP tasks
  - Syntactic parsing
  - Semantic role labeling
  - Semantic parsing
- Does WSD actually improve performance on some real end-user task?
  - Information retrieval
  - Information extraction
  - Machine translation
  - Question answering