Word Sense Disambiguation

WSD

- Word Sense Disambiguation:
  - Determine from context (or otherwise) what sense is intended a particular word
  - Useful, in particular, in
    - Query and Search
    - IR-related tasks
  - Several types of methods for WSD:
    - Dictionary-based
    - Supervised
    - Unsupervised

Lexical Ambiguity

- *All* words have multiple senses
- Much of the CL literature not completely clear on what is meant by a word sense
- Take the word *bank*
  1. Financial Institution sense
  2. Sloping land, as in "river bank"
- These senses obviously semantically quite far apart

Lexical Ambiguity

- Additional senses of *bank* from Wordnet:
  3. a building in which the business of banking transacted
  4. an arrangement of similar objects in a row or in tiers
  5. a container for keeping money at home
  6. a long ridge or pile
  7. the funds held by a gambling house or the dealer in some gambling games
  8. a slope in the turn of a road or track
  9. a flight maneuver; aircraft tips laterally about its longitudinal axis
  10. a supply or stock held in reserve for future use (especially in emergencies)

Homophony vs. Polysemy

- Senses 1 & 2 are homophones (and homonyms and homographs)
- Homophones are unrelated senses, resulting (probably) from some historical convergence onto a form
- (Or historic divergence of meaning.)
- Polysemes are often metaphorical, or constitute semantically very similar notions to some primary sense
- Wordnet makes no distinction!
- Some dictionaries do (e.g. Webster's), others do not (e.g., Longman's)

Homophony vs. Polysemy

- Why does it matter?
- Imagine a query:
  - User requests "banks Puget Sound"
  - Query engine may return financial institutions in the Seattle area (or blood banks)
- An error between sense 1 & 2 a bigger error than between 1 and its related senses

Differentiating Senses

- A thesaurus or semantic network can help tease apart the senses
- Roget's electronic thesaurus and WordNet are often used
- Differentiating senses involves looking at context
- Context: the words used in the "environment" of the target word (same sentence, same document, etc.)
- Mark the word with the most appropriate sense for the environment

WordNet

- A large semantic network (about 90,000 words, 200,000 senses, 600,000 links)
- Organized into different *synonym sets*, or *synsets*, by major categories (nouns, verbs, adjectives)
- Synsets tied together by different relations, such as hypernymy, hyponymy, antonymy, meronymy, etc.
- Hypernymy, hyponymy the most frequent relations