# Natural Language Processing

## Lecture #7
## Semantic Analysis I

Hutchatai Chanlekha


Department of Computer Engineering, Kasetsart University

# Lexical Semantics

- Semantic is the study of the meaning of linguistic utterances.
- Lexicon has systematic structure that governs what words can mean and how they can be used.
- Structure
  - Relations among words and their meanings
  - Internal structure of individual words
- **Lexical Semantics**: The study of this systematic structure
- Today class…
  - Focus on computational resources that capture lexical semantic information in a form that is useful for a wide variety of applications
- Term
  - **Lexeme**: individual entry in the lexicon
    - Pairing of a particular orthographic and phonological form with some form of symbolic meaning representation (i.e. sense)
  - **Lexicon**: finite list of lexemes

# Dictionary

- Dictionary
  - Repositories of information about the meanings of lexemes
- Consider the following fragment from dictionary

| | | |
|---|---|---|
| ***right*** | adj. | Located nearer the right hand esp. being on the right when facing the same direction as the observer |
| ***left*** | adj. | Located nearer to this side of the body than the right |
| ***red*** | n. | the color of blood or a ruby |
| ***blood*** | n. | the red liquid that circulates in the heart, arteries and veins of animals |

  - Circularity in the definitions …
- This circularity is evidence that dictionaries entries are rather descriptions of lexemes in terms of other lexemes

# Comparing ….

- http://www.ldoceonline.com/
  - http://www.ldoceonline.com/dictionary/withdraw
- http://dictionary.cambridge.org/
  - http://dictionary.cambridge.org/dictionary/british/withdraw
- http://dictionary.reference.com/
  - http://dictionary.reference.com/browse/withdraw?s=t
- http://www.oxforddictionaries.com/
  - http://www.oxforddictionaries.com/definition/english/withdraw

# Deciding word sense

- Lexicographers do not always agree on how to split a dictionary entry to senses.
  - Dictionaries often disagree with one another, as can be seen by comparing a pair of randomly selected dictionaries
  - It is not easy to decide when to lump two sense into one or when to split one sense into two.
    - Meaning is probably best thought of as a continuous quantity, with infinite number of shades between any two points.
    - Trade-off between lumping and splitting is often fairly arbitrary.

# Splitting entry in dictionary

- Dictionaries may split an entry when there are differences in:
  - **POS** (common)
  - **Syntactic features** (such as count/un-count nouns, person, number, gender, etc.)
  - **Valency structures** (e.g., transitive vs. intransitive verbs)
  - **Pronunciation** (rare and usually not the only reason for splitting senses)
  - **Etymology** (rare, especially in learners' dictionaries; more common in dictionaries based on historical principles.)
  - **Capitalization** (e.g., He = "god", East = "(formerly) Communist Countries")
  - **Register** (e.g., rude, slang)
  - **Dialect** (e.g., US, British, Canadian)
  - **Collocations, phrases** (e.g., eat away at, eat in, eat up)
  - **Subject codes** (subject codes are usually not given in the written text but they can be found in the electronics versions of a few dictionaries)

From: Gale, W.A., Church, K. W., and Yarowsky D., 1993, "A Method for Disambiguating Word Senses in a Large Corpus"

# Relations Among Lexemes and Their Sense

We can capture quite a bit about semantics of individual lexemes by analyzing and labeling their relations to other lexemes in various settings.

- Homonymy
- Polysemy
- Synonymy ↔ Antonym
- Hyponymy ↔ Hypernym

# Homonymy

- Relation that holds between words that have the same form with unrelated meanings
- **Homonyms**: Items taking part in homonymy relation
- Example:
  - Bank
    - Financial institution VS sloping mound
  - ขัน
    - ขันน้ำ vs ขำขัน vs ไก่ขัน VS ขันน็อต

# Homonymy (cont.)

- Homonymy
  - Normally only dictionary entries with **identical citation-forms** are considered candidates for homonymy
  - Citation forms are orthographic forms that are used to alphabetically index words in a dictionary
- Words with the same pronunciation but different spellings
  - Not considered as homonyms, but rather *homophones*
    - be VS bee // สาด VS ศาสตร์
- Words with identical orthographic forms but different pronunciations
  - Rarely appear in traditional list of homonyms
  - They are usually called *homographs*
    - dessert VS dessert // สระ VS สระ
- Lexemes with different POS are also typically not considered to be good candidates for homonymy.

# Homonymy (cont.)

- Difficulties in applications
  - Speech recognition
    - Homophones, such as *to*, *two*, *too*, cause obvious problems
  - ML model
    - For perfect homonyms, the entries for all the distinct lexemes are conflated, which results in inappropriate probability assignment
      - Example: Suppose $w_1$ and $w_2$ are homonym

        $w_1$ and $a$ always occur together → $w_1\ a$

        $w_2$ occurs with $b$, $c$, $d$ → $w_2\ b$, $w_2\ c$, $w_2\ d$

        - Homonym reduce probability assigned to $w_1\ a$
  - Text-to-Speech
    - Vulnerable to homographs with distinct pronunciations
  - Information retrieval
    - Performance degradation in the presence of homonyms/homographs

# Polysemy

- Polysemy
  - Multiple related meanings within a single lexeme
  - Example:
    - "*blood bank*" → clearly, "bank" does not refer to financial constitution
      → senses related to repositories for biological entities
  - Polysemy allows us to state that this sense of *bank* is related to, and possibly derived from, the financial institution sense, without asserting that it is a distinct lexeme
- Distinguishing homonymy from polysymy is not quite straightforward
  - Two criteria for determining whether the meanings of two lexemes are related or not
    - History (or etymology) of the lexemes in question
    - How lexemes are conceived of by native speakers
      - Coincidence: homonymy can be understood as accidentally share the same form, but not for polysemy

# How many sense??

- How to decide how many sense should be associated
  - Traditional approach
    - Lexicographers create entries with as many sense as necessary to account for all the distinctions in meaning observed in very large corpus
  - But….
    - Too much distinctions that are normally required for reasonable computational application
  - Example: Distinguish distinct senses
    - *They rarely serve red meat, preferring to prepare seafood.*
    - *He served as U.S. ambassador to Norway in 1976 and 1977*
    - *He might have served his time, come out and led an upstanding life.*
  - From example
    - Ex (1) … strong connection between  this sense and the notion of food preparation
    - Ex (2) … different syntactic subcategorization since its first argument is a PP.
      - Differing syntactic behaviors are often symptomatic of differing senses
    - Ex (3) … Incarceration, clear meaning without specific information about prison

# How many sense?? (cont.)

- Determine if two distinct sense are present ..
  - Combine two separate uses of a lexeme into a single example using a conjunction
  - Example:
    - “Which of those flights serve breakfast?”
    - “Does Midwest Express serve Philadelphia?”
    - ….. “Does Midwest Express serve breakfast and Philadelphia?” ……

# Discovering sense VS Determining sense

- Discovering the proper set of sense for a given lexeme is distinct from the process of determining which sense of a lexeme is being used in a given context
  - The latter task is called “Word sense disambiguation”
  - Word sense disambiguation
    - Presumes fixed set of sense for each lexical term
    - Identifies a word form in context with one sense chosen from that predetermined set of sense

# Synonymy

- Synonymy
  - Different lexemes with the same meaning
  - Simple test for synonymy
    - Notion of substitutability
      - Two lexemes will be considered synonyms if they can be substituted for one another in a sentence without changing either the meaning or acceptability of the sentence
      - Not necessary to be able to substitute in every sentence, just in some environment
  - Example:
    - *big* and *large*
    - *sure* and *certain*

# Synonymy (cont.)

- Failure of the substitution between synonyms
  - Polysemy
    - "Big sister" VS "Large sister"
    - "*big*" has as one of its distinct polysemous senses the notion of being older, or grown up, while "*large*" lack this sense
  - Subtle shades of meaning
    - Two lexemes share a central core meaning, but additional ancillary facts are associated with one of them
    - "*price*" and "*fare*"
      - *fare* → suited to the costs for services
  - Collocational constraints
    - "*big mistake*" VS "*large mistake*"
    - *"big"* and *"large"* have the same sense, but prefer "*big*" when using with "*mistake*"
  - Register
    - Lexemes with identical meaning, but are not interchangeable in some environments due to social factors such as politeness, group status, etc.

# Hyponymy

- Hyponymy
  - Pairings where one lexeme denotes a subclass of the other
  - Example: "*car*" and "*vehicle*"
  - Non-symmetric relation
    - More specific lexeme → hyponym
    - More generic lexeme → hypernym
    - Example: *car* is a hyponym of *vehicle* / *vehicle* is hypernym of *car*
  - Test of hyponymy relation
    - That is a *x* $\Rightarrow$ That is a *y*
      - If *x* is a hyponym of *y*, then if the sentence of the left is true then the sentence on the right must also be true.
    - That is a car $\Rightarrow$ That is a vehicle

# Effects of Homonymy, Polysemy, Synonymy on IR

- Problem of Homonymy, Polysemy
  - ***Reducing precision***
  - Example: $bank_1$ → financial institution สระ$_1$ → สระน้ำ
    $bank_2$ → sloping mound สระ$_2$ → ตัวอักษร
  - Query containing "bank" will be judged similar to documents making use of either of these senses
  - If user wants $bank_2$, documents containing $bank_1$ will be judged irrelevant

- Lack of synonymy and hyponymy
  - ***Reducing recall***
  - Example: query term ➔ "dog"
    - Match documents that make frequent use of "dog"
    - But fail to match documents that use close synonyms like "canine" or documents that use hyponyms like "Malamute"

# WordNet

- Lexical database for English
  - Lexical and lexical relation
- http://wordnet.princeton.edu/

| Relation | Definition | Example |
|---|---|---|
| Hypernym | From concepts to superordinates | breakfast → meal |
| Hyponym | From concepts to subtypes | meal → lunch |
| Has-Member | From groups to their members | faculty → professor |
| Member-Of | From members to their groups | copilot → crew |
| Has-Part | From wholes to parts | table → leg |
| Part-Of | From parts to wholes | course → meal |
| Antonym | Opposites | leader ↔ follower |

Noun relations in WordNet

| Relation | Definition | Example |
|---|---|---|
| Antonym | Opposite | heavy ↔ light<br>quickly ↔ slowly |

Adjective and adverb relations in WordNet

| Relation | Definition | Example |
|---|---|---|
| Hypernym | From concepts to super-ordinates | fly → travel |
| Troponym | From events to their subtypes | sleep → nap |
| Entails | From events to the events they entail | snore → sleep |
| Antonym | Opposites | increase ↔ decrease |

Verb relations in WordNet

Synonymy in WordNet is organized around the notion of a **synset**, a set of synonyms

# Word Sense Disambiguation

# Word sense disambiguation

- Word sense disambiguation (WSD) is defined as the task of assigning the appropriate meaning (sense) to a given word in a text or discourse, when the **word** has multiple meanings.
  - Example:
    - bass

      (a) I can hear bass sounds.

      (b) They like grilled bass.
    - bank

      (a) The river bank was full of dead fishes.

      (b) I went to the bank to deposit my money

# WSD approach

- Survey of word sense disambiguation approach
  - http://www.cse.unt.edu/~tarau/teaching/NLP/papers/ACM_Survey_2009_Navigli.pdf
  - https://arxiv.org/abs/1508.01346