# CMPT-413
# Computational Linguistics

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

# Lexical Semantics

- So far, we have listed **words** in our **lexicon** or **vocabulary** assuming a single meaning per word:
  Consider $n$-grams $P(w_i \mid w_{i-2}, w_{i-1})$ = *P(Bank | on, Commerce)* or prepositional phrase attachment *if p=on and n2=bank then change N to V*

- Consider . . . *withdraw twenty dollars on the bank* (correct = $V$) vs.
  . . . *withdraw the troops on the bank* (correct = $N$)

- The same word *bank* means two different things but we cannot distinguish between them using the traditional definition of word.

# Lexical Semantics

- To deal with this issue, we combine the *spelling* or *pronunciation* of a word and the *meaning*.
  In the *lexicon* we now store **lexemes** instead of words. A lexeme pairs a particular spelling or pronunciation with a particular meaning.

- The meaning part of a lexeme is called a **sense**. For CL, our interest is in relations between lexemes or disambiguating different senses of a word.
  word: bank → lexeme: **bank**$^{1}$ *OR* word: bank → lexeme: **bank**$^{2}$

- Note that meanings are often not definitions, but often are simple listings of compatible lexemes.
  cf. dictionary defns: *red, n.* the color of blood or ruby; *blood, n.* red liquid circulating in animals

# Homonyms

- Homonyms: *words that have the same form but different meanings*

  1. *Instead, the chemical* plant *was found in violation of several environmental laws*

  2. *Stanley formed an expedition to find a rare* plant *found along the Amazon river*

- Same *orthographic* form: *plant* but two senses: **plant**$^{1}$ and **plant**$^{2}$

# Homonyms

- Text vs. speech: fly-casting for *bass* vs. rhythmic *bass* chords
  These cases are homonyms in text, but not in speech. Referred to as **homographs**

- Speech vs. text: *would* vs. *wood*
  These cases are not homonyms in text, but easily confused in speech. Referred to as **homophones**

- Note that this problem in some cases can be solved using *part of speech tagging*
  Can you think of a case which cannot be solved using POS tagging?

# Applications

- Spelling correction: homophones: *weather* vs. *whether*
- Speech recognition: homophones: *to*, *two*, *too*. Also homonyms (see $n$-gram e.g.)
- Text to speech: homographs: *bass* vs. *bass*
- Information retrieval: homonyms: *latex*

# Polysemy

- Consider the homonym: *bank* → commercial **bank**$^1$ vs. river **bank**$^2$

- Now consider

  1. *A PCFG can be trained using derivation trees from a tree* bank *annotated by human experts*

- Is this a new sense of *bank*?

# Polysemy

- Senses can be derived from a particular lexeme. This process is known as **polysemy**
  In previous case we would say that the use of *bank* is a sense derived from commercial **bank**$^{1}$

- In some cases, splitting into different lexemes has other supporting evidence: **bank**$^{1}$ has Italian origin vs. **bank**$^{2}$ has Scandinavian origin

  1. *A PCFG can be trained using a* bank *of derivation trees called a tree-bank annotated by human experts*

- How can we tell between homonyms and polysemous uses of a word?

# Zeugma

- Consider the case for a verb like *serve*

1. *Does United serve breakfast?*
2. *Does United serve Philadelphia?*
3. *Does United serve breakfast and dinner?*
4. *#Does United serve breakfast and Philadelphia?*

# Word Sense Disambiguation

- Consider a noun like *bank*

  1. *How many senses does it have?*

  2. *How are these senses related?*

  3. *How can they be reliably distinguished?*

- For NLP software, among these three questions, typically at runtime we need to automatically find the answer to the last question: given a word in context, map it to the correct lexeme: **word-sense disambiguation**

# Word Sense Disambiguation: training data

```
 training_VBG new_JJ Ukrainian_JJ    plant(1)    _NN operators_NNS to_TO replace_V
who_WP are_VBP leaving_VBG the_DT    plant(1)    s_NNS in_IN Ukraine_NNP and_CC im
CC safety_NN procedures_NNS at_IN    plant(1)    s_NNS in_IN both_DT countries_NNS
t_IN the_DT Orange_NNP County_NNP    plant(1)    _NN ._.
Z closing_VBG three_CD missile_NN    plant(1)    s_NNS in_IN southern_JJ Californi
_IN the_DT whole_JJ Chernobyl_NNP    plant(1)    _NN in_IN 1991_CD ,_, five_CD yea
IN a_DT hill_NN ,_, gardeners_NNS    plant(2)    _NN begonias_NNS ,_, making_VBG f
$_$ 200_CD million_CD printing_NN    plant(1)    _NN in_IN Brooklyn_NNP ,_, Ohio_N
of_IN incompletely_JJ oxidated_JJ    plant(2)    _NN and_CC animal_NN sediment_NN
whenever_WRB you_PRP eat_VBP a_DT    plant(2)    _NN ._. ''_''
n_IN return_NN for_IN a_DT new_JJ    plant(1)    _NN near_IN Tuscaloosa_NNP ._.
T carmaker_NN could_MD finance_VB    plant(1)    _NN construction_NN with_IN the_D
n_IN return_NN for_IN a_DT new_JJ    plant(1)    _NN near_IN Tuscaloosa_NNP ._.
```

# Word Sense Disambiguation: learning

- Many different learning methods: let's consider one, Transformation Based Learning

- Let rule condition
  $r \leftarrow \mathtt{W}_{-1} = \mathtt{gardeners}, \mathtt{W}_{+1} = \mathtt{begonias}, \mathtt{W}_{+\mathtt{window}} = \mathtt{floral}$

- If $r$ then change from **plant**$^1$ (manufacturing plant) to **plant**$^2$ (living plant)

# Synonyms

- Synonyms: Different lexemes with the same meaning

  1. *How big/large is that plane?*

  2. *Would I be flying on a big/large or small plane?*

- Synonyms clash with polysemous meanings

  1. *Seema is my big sister*

  2. *#Seema is my large sister*

# WordNet

- `WordNet` is an electronic database of word relationships, handcrafted from scratch by researchers at Princeton University (George Miller, Christine Fellbaum, et al.)

- WordNet contains 3 databases: for verbs, nouns and one for adjectives and adverbs

| Category | Unique Forms | Number of Senses |
|---|---|---|
| Noun | 94474 | 116317 |
| Verb | 10319 | 22066 |
| Adjective | 20170 | 29881 |
| Adverb | 4546 | 5677 |

# WordNet

- Ask the question: how many senses per noun or verb? The distribution of senses follows Zipf's (2nd) Law.

- WordNet provides multiple lexeme entries for each word and for each part of speech,
  e.g. *plant* as noun has 3 senses; *plant* as verb has 2 senses

- WordNet also provides *domain-independent* lexical relations such as IS-A, HasMember, MemberOf, . . .

# WordNet: noun relations

| Relation | Definition | Example |
|---|---|---|
| Hypernym | this is a kind of | *breakfast* → *meal* |
| Hyponym | this has a specific instance | *meal* → *lunch* |
| Has-Member | this has a member | *faculty* → *professor* |
| Member-Of | this is member of a group | *copilot* → *crew* |
| Has-Part | this has a part | *table* → *leg* |
| Part-Of | this is part of | *course* → *meal* |
| Antonym | this is an opposite of | *leader* → *follower* |

# WordNet: verb relations

| Relation | Definition | Example |
|---|---|---|
| Hypernym | this event is a kind of | *fly* → *travel* |
| Tropynym | this event has a subtype | *walk* → *stroll* |
| Entails | this event entails | *snore* → *sleep* |
| Antonym | this event is opposite of | *increase* → *decrease* |

# WordNet: example from ver1.7.1

Sense1: Canada

⇒North American country,North American nation

⇒country, state, land

⇒administrative district,administrative division,territorial division

⇒district, territory

⇒region

⇒location

⇒entity, physical thing

# WordNet: example from ver1.7.1

Sense 3: Vancouver
- ⇒city, metropolis, urban center
  - ⇒municipality
    - ⇒urban area
      - ⇒geographical area
        - ⇒region
          - ⇒location
            - ⇒entity, physical thing
    - ⇒administrative district, territorial division
      - ⇒district, territory
        - ⇒region
          - ⇒location
            - ⇒entity, physical thing
- ⇒port
  - ⇒geographic point
    - ⇒point
      - ⇒location
        - ⇒entity, physical thing

# WordNet

- A **synset** in WordNet is a list of synonyms (interchangeable words)

- { chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug }

- How can we use this information like synsets, hypernyms, etc. from WordNet to benefit NLP applications?

- Consider one example: PP attachment, words plus word classes extracted from the hypernym hierarchy increase accuracy from 84% to 88% (Stetina and Nagao, 1998)

# WordNet

- Another example of WordNet used in NLP applications: **selectional restrictions**

- We have considered subcategorization:
  *VP-with-NP-complement* → *V(eat) NP* "eat six bowls of rice "
  But not selectional restrictions of the verb itself: " *eat tomorrow* "
  Consider *what do you want to eat tomorrow*

- We can use the **synset** { `food, nutrient` } to describe the NP argument of *eat* – then the 60K lexemes under these nodes in the WordNet hierarchy will be acceptable.
  (however, what about " *eat my shorts* ")
  → several other applications have been explored