Meaning: Lexical Semantics

Instructor: Jackie CK Cheung and David Adelani

COMP-550

J&M Ch. 16, 17–17.3 (1st ed); J&M Ch. 19, 20–20.5 (2nd ed); J&M Ch. 6.1; 23.2 – 23.5 (3rd ed older version)

Review Quiz

What are the components of a PCFG?

Vanilla PCFGs

Estimate of rule probabilities:

- MLE estimates:

$$\Pr(\alpha \rightarrow \beta) = \frac{\#(\alpha \rightarrow \beta)}{\#\alpha}$$

- e.g., Pr(S -> NP VP) = #(S -> NP VP) / #(S)
  - Recall: these distributions are normalized by LHS symbol

Even with smoothing, doesn’t work very well:

- Not enough context
- Rules are too sparse

Subject vs Object NPs

NPs in subject and object positions are not identically distributed:

- Obvious cases – pronouns (*I* vs *me*)
  - But both appear as NP -> PRP -> *I/me*
- Less obvious: certain classes of nouns are more likely to appear in subject than object position, and vice versa.
  - For example, subjects tend to be **animate** (usually, humans, animals, other moving objects)

Many other cases of obvious dependencies between distant parts of the syntactic tree.

Sparsity

Consider subcategorization of verbs, with modifiers

• *ate*	VP -> VBD
• *ate quickly*	VP -> VBD AdvP
• *ate with a fork*	VP -> VBD PP
• *ate a sandwich*	VP -> VBD NP
• *ate a sandwich quickly*	VP -> VBD NP AdvP
• *ate a sandwich with a fork*	VP -> VBD NP PP
• *quickly ate a sandwich with a fork*	VP -> AdvP VBD NP PP

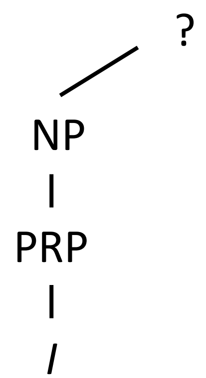
We should be able to factorize the probabilities:

- of having an adverbial modifier, of having a PP modifier, etc.

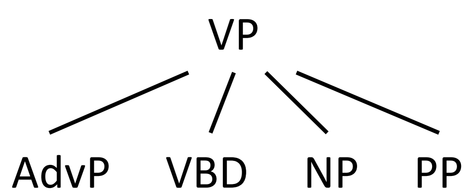
Wrong Independence Assumptions

Vanilla PCFGs make independence assumptions that are too strong AND too weak.

Too strong: *vertically*, up and down the syntax tree



Too weak: *horizontally*, across the RHS of a production



Adding Context

Add more context vertically to the PCFG

- Annotate with the parent category

Before: NP -> PRP, NP -> Det NN, etc.

Now:

Subjects:

NP^S -> PRP, NP^S -> Det NN, etc.

Objects:

NP^VP -> PRP, NP^VP -> Det NN, etc.

Learn the probabilities of the rules separately (though they may influence each other through interpolation/smoothing)

Example

Let's help Pierre Vinken find his ancestors.

```
( (S
    (NP
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP (NNP Nov.) (CD 29) )))
    (. .) ))
```

Note that the tree here is given in bracket parse format, rather than drawn out as a graph.

Removing Context

Conversely, we break down the RHS of the rule when estimating its probability.

Before: Pr(VP -> START AdvP VBD NP PP END) as a unit

Now: Pr(VP -> START AdvP) *

Pr(VP -> AdvP VBD) *

Pr(VP -> VBD NP) *

Pr(VP -> NP PP) *

Pr(VP -> PP END)

- In other words, we're making the same N-gram assumption as in language modelling, only over non-terminal categories rather than words.
- Learn probability of factors separately

Example

Let’s help Pierre Vinken find his children.

```
( (S
    (NP
      (NP (NNP Pierre) (NNP Vinken) )
      (, ,)
      (ADJP
        (NP (CD 61) (NNS years) )
        (JJ old) )
      (, ,) )
    (VP (MD will)
      (VP (VB join)
        (NP (DT the) (NN board) )
        (PP (IN as)
          (NP (DT a) (JJ nonexecutive) (NN director) ))
        (NP (NNP Nov.) (CD 29) )))
    (. .) ))
```

Markovization

Vertical markovization: adding ancestors as context

- Zeroth order – vanilla PCFGs
- First order – the scheme we just described
- Can go further:
  - e.g., Second order: NP^VP^S -> …

Horizontal markovization: breaking RHS into parts

- Infinite order – vanilla PCFGs
- First order – the scheme we just described
- Can choose any other order, do interpolation, etc.

Effect of Category Splitting

Vertical Order		Horizontal Markov Order $h = 0$	$h = 1$	$h \leq 2$	$h = 2$	$h = \infty$
$v = 1$	No annotation	71.27 (854)	72.5 (3119)	73.46 (3863)	72.96 (6207)	72.62 (9657)
$v \leq 2$	Sel. Parents	74.75 (2285)	77.42 (6564)	77.77 (7619)	77.50 (11398)	76.91 (14247)
$v = 2$	All Parents	74.68 (2984)	77.42 (7312)	77.81 (8367)	77.50 (12132)	76.81 (14666)
$v \leq 3$	Sel. GParents	76.50 (4943)	78.59 (12374)	79.07 (13627)	78.97 (19545)	78.54 (20123)
$v = 3$	All GParents	76.74 (7797)	79.18 (15740)	79.74 (16994)	79.07 (22886)	78.72 (22002)

Figure 2: Markovizations: F_1 and grammar size.

WSJ results by Klein and Manning (2003)

- With additional linguistic insights, they got up to 87.04 F1
- Current best is around 94-95 F1

Where Are We In the Course?

Single decisions	→	Text classification
Sequences	→	Language modelling
		Sequence labelling
Structure	→	Parsing

Next big topic: **semantics**

Semantics

The study of **meaning** in language

What does meaning mean?

- Relationship of linguistic expression to the real world
- Relationship of linguistic expressions to each other

Let's start by focusing on the meaning of **words**—**lexical semantics**.

Later on:

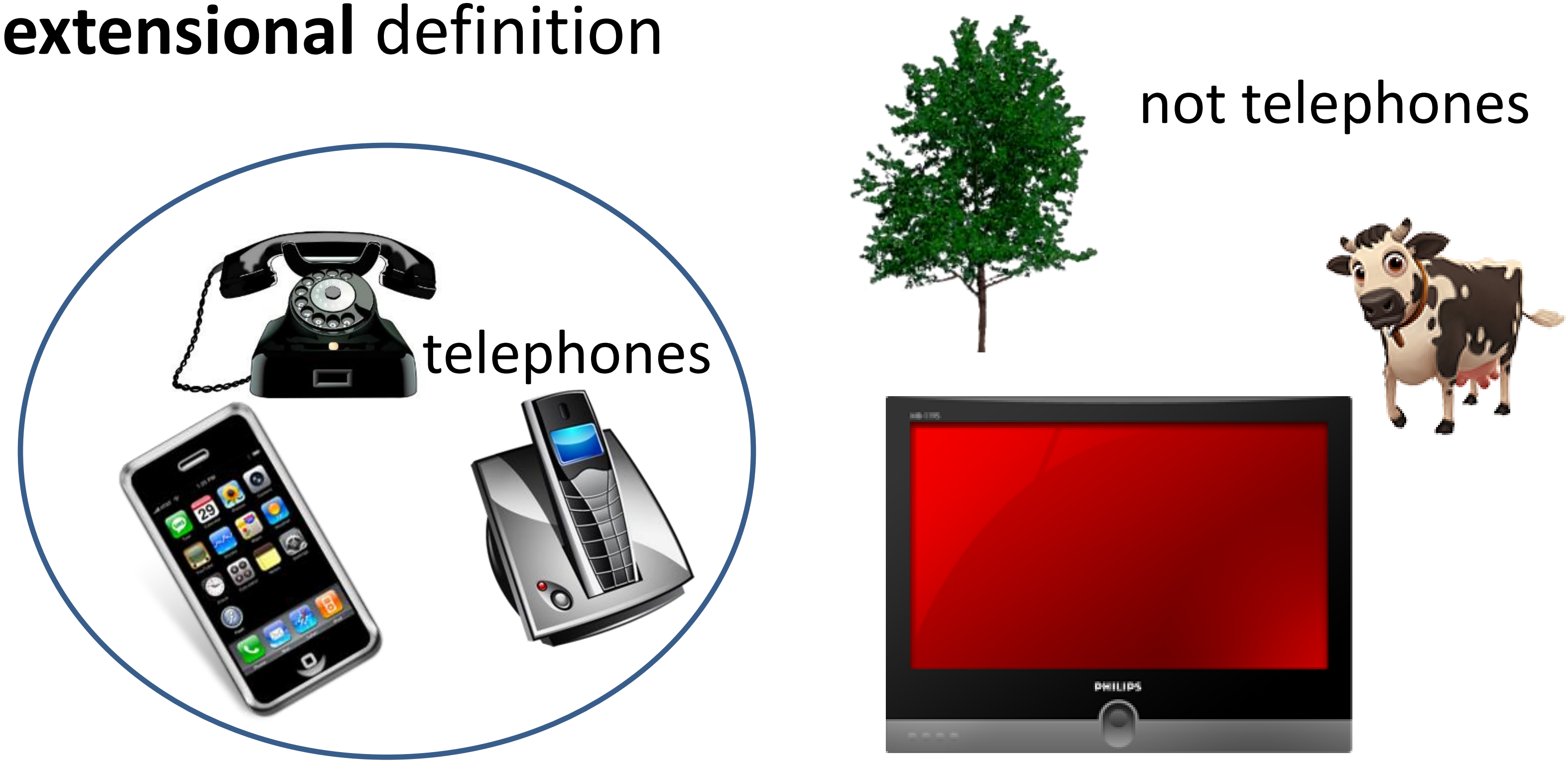
- meaning of phrases and sentences
- how to construct that from meanings of words

From Language to the World

What does *telephone* mean?

- Picks out all of the objects in the world that are telephones (its **referents**)

Its **extensional** definition



Relationship of Linguistic Expressions

How would you define *telephone*? e.g, to a three-year-old, or to a friendly Martian.

Dictionary Definition

http://dictionary.reference.com/browse/telephone

Its **intensional definition**

- The necessary and sufficient conditions to be a telephone

This presupposes you know what “apparatus”, “sound”, “speech”, etc. mean.

Sense and Reference (Frege, 1892)

Frege was one of the first to distinguish between the **sense** of a term, and its **reference.**

Same referent, different senses:

Venus →

the morning star →

the evening star →

Lexical Semantic Relations

How specifically do terms relate to each other? Here are some ways:

- Hypernymy/hyponymy
- Synonymy
- Antonymy
- Homonymy
- Polysemy
- Metonymy
- Synecdoche
- Holonymy/meronymy

Hypernymy/Hyponymy

ISA relationship

Hyponym	Hypernym
monkey	*mammal*
Montreal	*city*
red wine	*beverage*

Synonymy and Antonymy

Synonymy

(Roughly) same meaning

offspring   descendent   spawn

happy   joyful   merry

Antonymy

(Roughly) opposite meaning

synonym   antonym

happy   sad

descendant ancestor

Homonymy

Same form, different (and unrelated) meaning

Homophone – same sound

- e.g., *son* vs. *sun*

Homograph – same written form

- e.g., *lead* (noun) vs. *lead* (verb)

Polysemy

Multiple related meanings

S: (n) **newspaper**, paper (a daily or weekly publication on folded sheets; contains news and articles and advertisements) *"he read his newspaper at breakfast"*

S: (n) **newspaper**, paper, newspaper publisher (a business firm that publishes newspapers) *"Murdoch owns many newspapers"*

S: (n) **newspaper**, paper (the physical object that is the product of a newspaper publisher) *"when it began to rain he covered his head with a newspaper"*

S: (n) **newspaper**, newsprint (cheap paper made from wood pulp and used for printing newspapers) *"they used bales of newspaper every day"*

Homonymy vs Polysemy

Homonymy: unrelated   Polysemy: related meaning

- S: (n) **position**, place (the particular portion of space occupied by something) *"he put the lamp back in its place"*
- S: (n) military position, **position** (a point occupied by troops for tactical reasons)
- S: (n) **position**, view, perspective (a way of regarding situations or topics etc.)*"consider what follows from the positivist view"*
- S: (n) **position**, posture, attitude (the arrangement of the body and its limbs) *"he assumed an attitude of surrender"*
- S: (n) status, **position** (the relative position or standing of things or especially persons in a society) *"he had the status of a minor"; "the novel attained the status of a classic"; "atheists do not enjoy a favorable position in American life"*
- S: (n) **position**, post, berth, office, spot, billet, place, situation (a job in an organization) *"he occupied a post in the treasury"*

Metonymy

Substitution of one entity for another related one

We ordered many delicious dishes at the restaurant.

I worked for the local paper for five years.

Quebec City is cutting our budget again.

The loonie is at a 11-year low.

Synecdoche – a specific kind of metonymy involving whole-part relations

All hands on deck!

Don’t be a <censored body part>

Holonymy/meronymy

Some kind of whole/part relationship

Subtypes	Holonym	Meronym
groups and members	*class*	*student*
whole and part	*car*	*windshield*
whole and substance	*chair*	*wood*

Quiz

Classify the following examples in terms of what lexical semantic relation they exhibit

cold	*freezing*
they're	*their*
hair	*head*
enemy	*friend*
cut (hair)	*cut (bread)*
George Clooney	*actor*

WordNet (Miller et et., 1990)

WordNet is a lexical resource organized by **synsets**

- Nodes: synsets
- Edges: lexical semantic relation between two synsets

Separate hierarchy for different parts of speech

- Nouns, verbs, adjectives, adverbs

WordNet online:

http://wordnetweb.princeton.edu/perl/webwn

A Synset Entry

S: (n) **hand**, manus, mitt, paw (the (prehensile) extremity of the superior limb) *"he had the hands of a surgeon"; "he extended his mitt"*

direct hyponym / *full hyponym*

- S: (n) fist, clenched fist (a hand with the fingers clenched in the palm (as for hitting))
- S: (n) hooks, meat hooks, maulers (large strong hand (as of a fighter))*"wait till I get my hooks on him"*
- S: (n) right, right hand (the hand that is on the right side of the body) *"he writes with his right hand but pitches with his left"; "hit him with quick rights to the body"*
- S: (n) left, left hand (the hand that is on the left side of the body) *"jab with your left"*

part meronym

direct hypernym / *inherited hypernym* / *sister term*

part holonym

- S: (n) arm (a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb)
- S: (n) homo, man, human being, human (any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage)

derivationally related form

http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=hand&i=8&h=11000000000000000000000#c

WordNet Has an NLTK Interface

>>> from nltk.corpus import wordnet

Some useful functions:

>>> wordnet.synsets(*<query_term>*)

>>> wordnet.synset(*<synset_name>*)

Remember you can use dir and help to get a list of functions in Python.

Word Sense Disambiguation

Figuring out which word sense is expressed in context

*His **hands** were tired from hours of typing.*

→ *hand.n.01*

*Due to her superior education, her **hand** was flowing and graceful.*

→ *hand.n.03*

General idea: use words in the context to disambiguate.

Which words above would help with this?

Possible Computational Approaches

A heuristic algorithm

- **Lesk's algorithm**

Supervised machine learning

- Possible, but requires a lot of work to annotate word sense information that we want to avoid

Unsupervised, or minimally supervised machine learning

- **Yarowsky's algorithm**

Lesk's Algorithm (1986)

Use the dictionary definitions of a word's senses

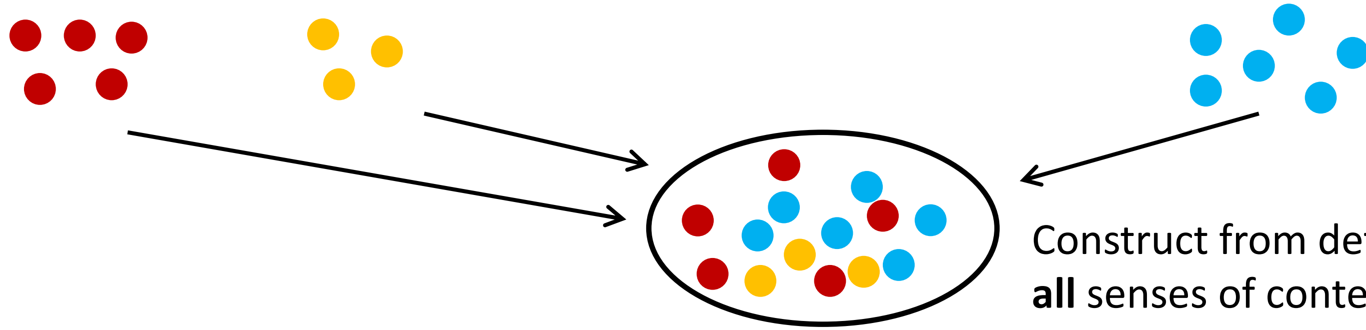
Steps to disambiguate word w:

1. Construct a bag of words representation of the context, B
2. For each candidate sense s_i of word w:
   - Calculate a signature of the sense by taking all of the words in the dictionary definition of s_i
   - Compute $\mathrm{Overlap}(B, \mathrm{signature}(s_i))$
3. Select the sense with the highest overlap score

Financial Bank or Riverbank?

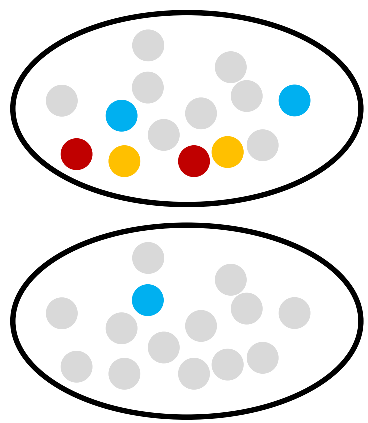
… ***deposit*** *a* ***cheque*** *at the* <u>*bank*</u> *before it* ***closed*** …

- B:



Construct from definitions of **all** senses of context words

- $overlap(bank\#1, B)$
  - 6 overlaps found
- $overlap(bank\#2, B)$
  - 1 overlap found
- Decision: select sense 1.



Model Variations

Which dictionary to use? NLTK?

Use only dictionary definitions? Or include example sentences?

Ignore uninformative stopwords (e.g., *the, a, of*)?

Lemmatize when considering matches (*tomatoes* matches *tomato*)?

Exercise

Run the Lesk algorithm using NLTK/WordNet. Ignore stop words, include examples, count lemma overlap. Consider only the top two senses of bank.

1. *I'll <u>deposit</u> the <u>cheque</u> at the **bank**.*
2. *The **bank** <u>overflowed</u> and <u>water</u> <u>flooded</u> the <u>town</u>.*

Yarowsky's Algorithm (1995)

A method based on **bootstrapping**

Steps:

1. Gather a data set with target word to be diambiguated
2. Automatically label a small **seed set** of examples
3. Repeat the following for a while:
    - Train a supervised learning algorithm from the seed set
    - Apply the supervised model to the entire data set
    - Keep the highly confident classification outputs to be the new seed set
4. Use the last model as the final model

Yarowsky's Example

Step 1: Disambiguating *plant*

Sense	Training Examples (Keyword in Context)
?	... company said the *plant* is still operating
?	Although thousands of *plant* and animal species
?	... zonal distribution of *plant* life . ...
?	... to strain microscopic *plant* life from the ...
?	vinyl chloride monomer *plant* , which is ...
?	and Golgi apparatus of *plant* and animal cells
?	... computer disk drive *plant* located in ...
?	... divide life into *plant* and animal kingdom
?	... close-up studies of *plant* life and natural
?	... Nissan car and truck *plant* in Japan is ...
?	... keep a manufacturing *plant* profitable without
?	... molecules found in *plant* and animal tissue
?	... union responses to *plant* closures . ...
?	... animal rather than *plant* tissues can be
?	... many dangers to *plant* and animal life
?	company manufacturing *plant* is in Orlando ...
?	... growth of aquatic *plant* life in water ...
?	automated manufacturing *plant* in Fremont ,
?	... Animal and *plant* life are delicately
?	discovered at a St. Louis *plant* manufacturing
?	computer manufacturing *plant* and adjacent ...
?	... the proliferation of *plant* and animal life
?	... ...

Step 2: Initial Seed Set

Sense A:

- *plant* as in a lifeform

Other data

Sense B:

- *plant* as in a factory

Sense	Training Examples (Keyword in Context)
A	used to strain microscopic *plant* **life** from the ...
A	... zonal distribution of *plant* **life** . ...
A	close-up studies of *plant* **life** and natural ...
A	too rapid growth of aquatic *plant* **life** in water ...
A	... the proliferation of *plant* and animal **life** ...
A	establishment phase of the *plant* virus **life** cycle ...
A	... that divide **life** into *plant* and animal kingdom
A	... many dangers to *plant* and animal **life** ...
A	mammals . Animal and *plant* **life** are delicately
A	beds too salty to support *plant* **life** . River ...
A	heavy seas, damage , and *plant* **life** growing on ...
A	... ...
?	... vinyl chloride monomer *plant* , which is ...
?	... molecules found in *plant* and animal tissue
?	... Nissan car and truck *plant* in Japan is ...
?	... and Golgi apparatus of *plant* and animal cells ...
?	... union responses to *plant* closures . ...
?	... ...
?	... ...
?	... cell types found in the *plant* kingdom are ...
?	... company said the *plant* is still operating ...
?	... Although thousands of *plant* and animal species
?	... animal rather than *plant* tissues can be ...
?	... computer disk drive *plant* located in ...
B	... ...
B	automated **manufacturing** *plant* in Fremont ...
B	... vast **manufacturing** *plant* and distribution ...
B	chemical **manufacturing** *plant* , producing viscose
B	... keep a **manufacturing** *plant* profitable without
B	computer **manufacturing** *plant* and adjacent ...
B	discovered at a St. Louis *plant* **manufacturing**
B	... copper **manufacturing** *plant* found that they
B	copper wire **manufacturing** *plant* , for example ...
B	's cement **manufacturing** *plant* in Alpena ...
B	polystyrene **manufacturing** *plant* at its Dow ...
B	company **manufacturing** *plant* is in Orlando ...

Step 3: Train a Classifier

He went with a **decision-list** classifier (we didn't cover this one in class)

Initial decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
8.10	*plant* **life**	$\Rightarrow$ A
7.58	**manufacturing** *plant*	$\Rightarrow$ B
7.39	**life** (within $\pm$2-10 words)	$\Rightarrow$ A
7.20	**manufacturing** (in $\pm$2-10 words)	$\Rightarrow$ B
6.27	animal (within $\pm$2-10 words)	$\Rightarrow$ A
4.70	equipment (within $\pm$2-10 words)	$\Rightarrow$ B
4.39	employee (within $\pm$2-10 words)	$\Rightarrow$ B
4.30	assembly *plant*	$\Rightarrow$ B
4.10	*plant* closure	$\Rightarrow$ B
3.52	*plant* species	$\Rightarrow$ A
3.48	automate (within $\pm$2-10 words)	$\Rightarrow$ B
3.45	microscopic *plant*	$\Rightarrow$ A
	...	

Note how new collocations are found for each sense

Step 3: Change Seed Set

Use only the cases where classifier is highly confident

Labeling previously untagged contexts
using the one-sense-per-discourse property

Change in tag	Disc. Numb.	Training Examples (from same discourse)
A → A	724	... the existence of *plant* and animal life ...
A → A	724	... classified as either *plant* or animal ...
? → A	724	Although bacterial and *plant* cells are enclosed
A → A	348	... the life of the *plant* , producing stem
A → A	348	... an aspect of *plant* life , for example
? → A	348	... tissues ; because *plant* egg cells have
? → A	348	photosynthesis, and so *plant* growth is attuned

Results

96% on binary word sense distinctions

Same result as with supervised methods, but with minimal amounts of annotation effort!