COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Words & their Meaning: Word Sense Disambiguation

CMSC 470

Marine Carpuat

Slides credit: Dan Jurafsky

Today: Word Meaning

2 core issues from an NLP perspective

- **Semantic similarity**: given two words, how similar are they in meaning?
- **Word sense disambiguation**: given a word that has more than one meaning, which one is used in a specific context?

“Big rig carrying fruit crashes on 210 Freeway, creates jam”

http://articles.latimes.com/2013/may/20/local/la-me-ln-big-rig-crash-20130520

How do we know that a word (lemma) has distinct senses?

- Linguists often design tests for this purpose
- e.g., **zeugma** combines distinct senses in an uncomfortable way

```
Which flight serves breakfast?

Which flights serve BWI?

*Which flights serve breakfast
and BWI?
```

Word Senses

- “Word sense” = distinct meaning of a word
- Same word, different senses
  - Homonyms (homonymy): unrelated senses; identical orthographic form is coincidental
    - E.g., financial bank vs. river bank
  - Polysemes (polysemy): related, but distinct senses
    - E.g., Financial bank vs. blood bank vs. tree bank
  - Metonyms (metonymy): “stand in”, technically, a sub-case of polysemy
    - E.g., use “Washington” in place of “the US government”
- Different word, same sense
  - Synonyms (synonymy)

WordNet:
a lexical database for English

https://wordnet.princeton.edu/

- Includes most English nouns, verbs, adjectives, adverbs
- Electronic format makes it amenable to automatic manipulation: used in many NLP applications
- “WordNets” generically refers to similar resources in other languages

Synonymy in WordNet

- WordNet is organized in terms of “synsets”
  - Unordered set of (roughly) synonymous “words” (or multi-word phrases)

- Each synset expresses a distinct meaning/concept

WordNet: Example

Noun

{pipe, tobacco pipe} (a tube with a small bowl at one end; used for smoking tobacco)

{pipe, pipage, piping} (a long tube made of metal or plastic that is used to carry water or oil or gas etc.)

{pipe, tube} (a hollow cylindrical shape)

{pipe} (a tubular wind instrument)

{organ pipe, pipe, pipework} (the flues and stops on a pipe organ)

Verb

{shriek, shrill, pipe up, pipe} (utter a shrill cry)

{pipe} (transport by pipeline) “pipe oil, water, and gas into the desert”

{pipe} (play on a pipe) “pipe a tune”

{pipe} (trim with piping) “pipe the skirt”

WordNet 3.0: Size

Part of speech	Word form	Synsets
Noun	117,798	82,115
Verb	11,529	13,767
Adjective	21,479	18,156
Adverb	4,481	3,621
Total	155,287	117,659

Different inventories can be used to define senses

WordNet Sense	Spanish Translation	Roget Category	Target Word in Context
$bass^4$	lubina	FISH/INSECT	…fish as Pacific salmon and striped **bass** and…
$bass^4$	lubina	FISH/INSECT	…produce filets of smoked **bass** or sturgeon…
$bass^7$	bajo	MUSIC	…exciting jazz **bass** player since Ray Brown…
$bass^7$	bajo	MUSIC	…play **bass** because he doesn't have to solo…

Different inventories do not always agree on sense distinctions
e.g., translation makes some distinctions but not others

Exercise: how many senses of “drive”?

1. *"Can you drive this four-wheel truck?"*
2. *"We drive to the university every morning"*
3. *"We drive the car to the garage"*
4. *"He drives me mad"*
5. *"She is driven by her passion"*
6. *"Drive a nail into the wall"*
7. *" She is driving away at her doctoral thesis"*
8. *"What are you driving at?"*
9. *"My new truck drives well"*
10. *"She drives for the taxi company in Newark"*
11. *"drive the cows into the barn"*
12. *"We drive the turnpike to work"*
13. *"drive a golf ball"*

Exercise: how many senses of “drive”?

1. *"Can you drive this four-wheel truck?"*
2. *"We drive to the university every morning"*
3. *"We drive the car to the garage"*
4. *"He drives me mad"*
5. *"She is driven by her passion"*
6. *"Drive a nail into the wall"*
7. *" She is driving away at her doctoral thesis"*
8. *"What are you driving at?"*
9. *"My new truck drives well"*
10. *"She drives for the taxi company in Newark"*
11. *"drive the cows into the barn"*
12. *"We drive the turnpike to work"*
13. *"drive a golf ball"*

13 distinct senses according to WordNet!

Exercise: how many senses of “drive”?

1. *"We drive to the university every morning"* (operate or control a vehicle)
2. *"We drive the car to the garage"* (cause someone or something to move by driving)
3. *"He drives me mad"* (force into or from an action or state, either physically or metaphorically)
4. *"She is driven by her passion"* (to compel or force or urge relentlessly or exert coercive pressure on, or motivate strongly)
5. *"Drive a nail into the wall"* (push, propel, or press with force)
6. *" She is driving away at her doctoral thesis"* (strive and make an effort to reach a goal)
7. *"What are you driving at?" (*move into a desired direction of discourse)
8. *"My new truck drives well"* (have certain properties when driven)
9. *"She drives for the taxi company in Newark"* (work as a driver)
10. *"drive the cows into the barn"* (urge forward)
11. *"We drive the turnpike to work"* (proceed along in a vehicle)
12. *"drive a golf ball"* (strike with a driver, as in teeing off)

13 distinct senses according to WordNet!

What can we do when humans who annotate senses disagree?

- Disagreement is inevitable when annotating based on human judgments
  - Even with trained annotators
  - There is no “ground truth”

- We cannot measure “correctness” of annotations directly
- Instead, we can measure reliability of annotation
  - Do human annotators make same decisions consistently?
  - Assumption: high reliability implies validity

Quantifying (dis)agreement between human annotators: Cohen's Kappa

- Measures agreement between two annotators while taking into account the possibility of chance agreement

Probability of actual agreement

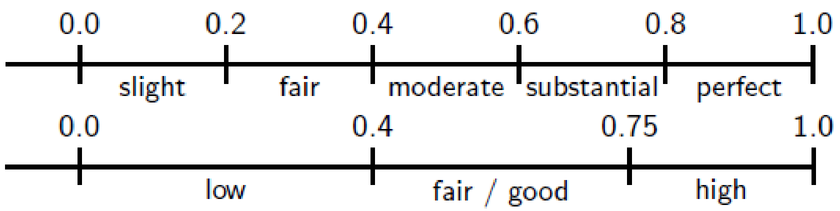
Probability of expected agreement

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

- Scales for interpreting Kappa

Landis & Koch, 1977

Green, 1997



Quantifying (dis)agreement between human annotators: Cohen's Kappa

Consider this confusion matrix for sense annotations by A and B of the same 250 examples

	Sense 1	Sense 2	Sense 3	Total
Sense 1	54	28	3	85
Sense 2	31	18	23	72
Sense 3	0	21	72	93
Total	85	67	98	250

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Here Pr(a) = 0.576, Pr(e) = 0.339, K=0.36
(agreement is low)

Word Sense Disambiguation what you should know (so far)

- Word senses distinguish different meanings of same word
- Sense inventories provide definitions of word senses
- Sense distinctions and annotations are based on human judgment
  - no “ground truth”
  - Measure annotation reliability using inter-annotator agreement

Word Sense Disambiguation

- Computational task
  - Given a predefined sense inventory (e.g., WordNet)
  - Goal: automatically select the correct sense of a word
  - Input: a word in context
  - Output: sense of the word

- Motivated by many applications:
  - Information retrieval
  - Machine translation
  - …

How hard is the problem?

- **Most words in English have only one sense**
  - 62% in Longman's Dictionary of Contemporary English
  - 79% in WordNet
- But the others tend to have several senses
  - Average of 3.83 in LDOCE
  - Average of 2.96 in WordNet

- **Ambiguous words are more frequently used**
  - In the British National Corpus, 84% of instances have more than one sense
- **Some senses are more frequent than others**

Baseline Performance

- Baseline: most frequent sense
  - Equivalent to “take first sense” in WordNet
  - Does surprisingly well!

Freq	Synset	Gloss
338	plant^1, works, industrial plant	buildings for carrying on industrial labor
207	plant^2, flora, plant life	a living organism lacking the power of locomotion
2	plant^3	something planted secretly for discovery by another
0	plant^4	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

62% accuracy in this case!

Upper Bound Performance

- Upper bound
  - Fine-grained WordNet sense: 75-80% human agreement
  - Coarser-grained inventories: 90% human agreement possible

Simplest WSD algorithm: Lesk's Algorithm

- Intuition: note word overlap between context and dictionary entries
  - **Unsupervised,** but knowledge rich

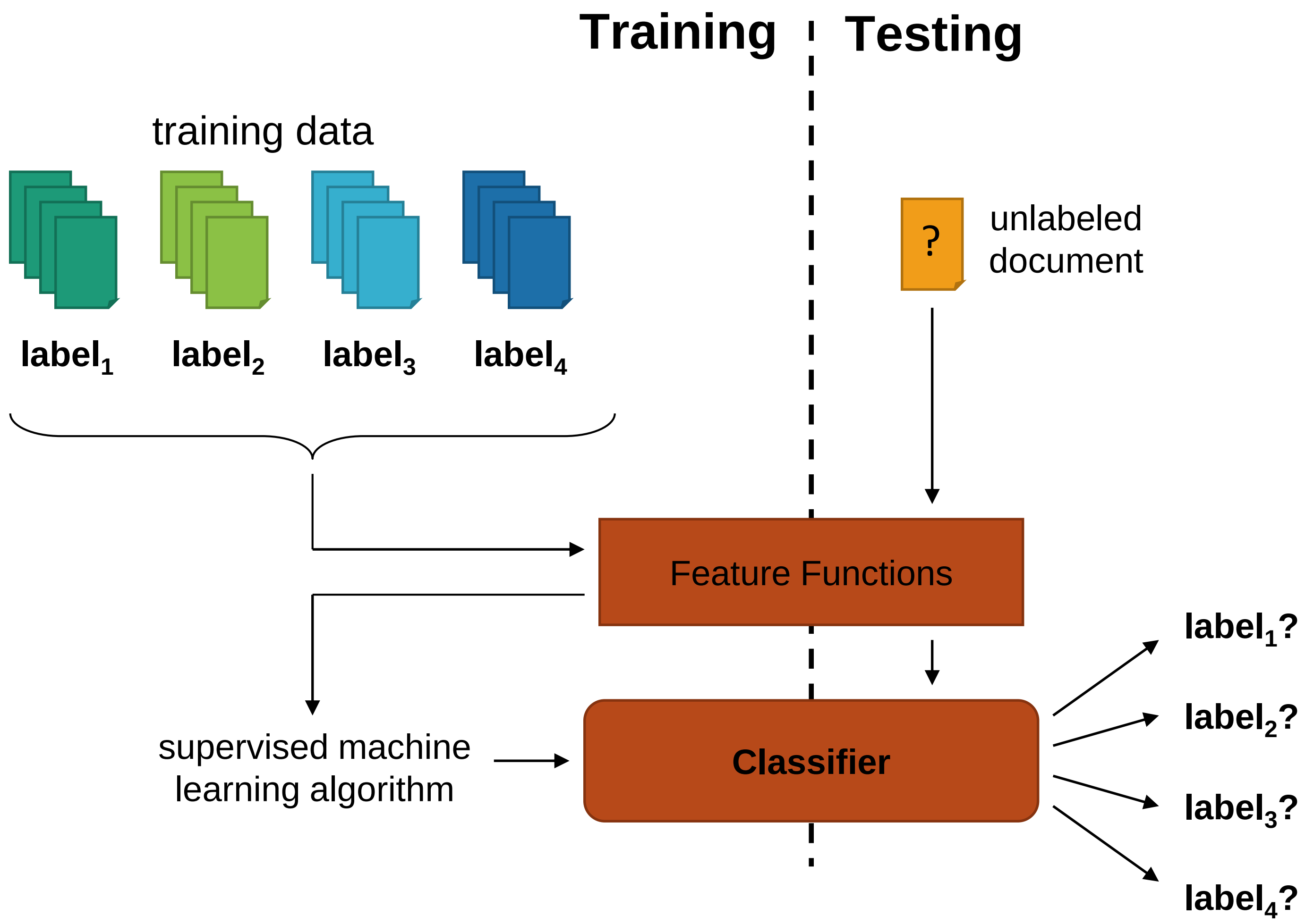
The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

Lesk's Algorithm

- Simplest implementation:
  - Count overlapping content words between glosses and context
- Lots of variants:
  - Include the examples in dictionary definitions
  - Include hypernyms and hyponyms
  - Give more weight to larger overlaps
  - Give extra weight to infrequent words (e.g., using idf)
  - …

Alternative: WSD as **Supervised** Classification



Existing Corpora

- Lexical sample
  - *line-hard-serve* corpus (4k sense-tagged examples)
  - *interest corpus* (2,369 sense-tagged examples)
  - …
- All-words
  - SemCor (234k words, subset of Brown Corpus)
  - Senseval/SemEval (2081 tagged content words from 5k total words)
  - …

How are annotated examples used in supervised learning?

- Supervised learning = requires examples annotated with correct prediction

- Used in 2 ways:
  - To find good values for the model (hyper)parameters (training data)
  - To evaluate how good the resulting classifier is (test data)

- How do we know how good a classifier is?
  - Compare classifier predictions with human annotation
  - On held out test examples
  - Evaluation metrics: accuracy, precision, recall

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Precision and recall

- **Precision**: % of selected items that are correct
  Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

A combined measure: F

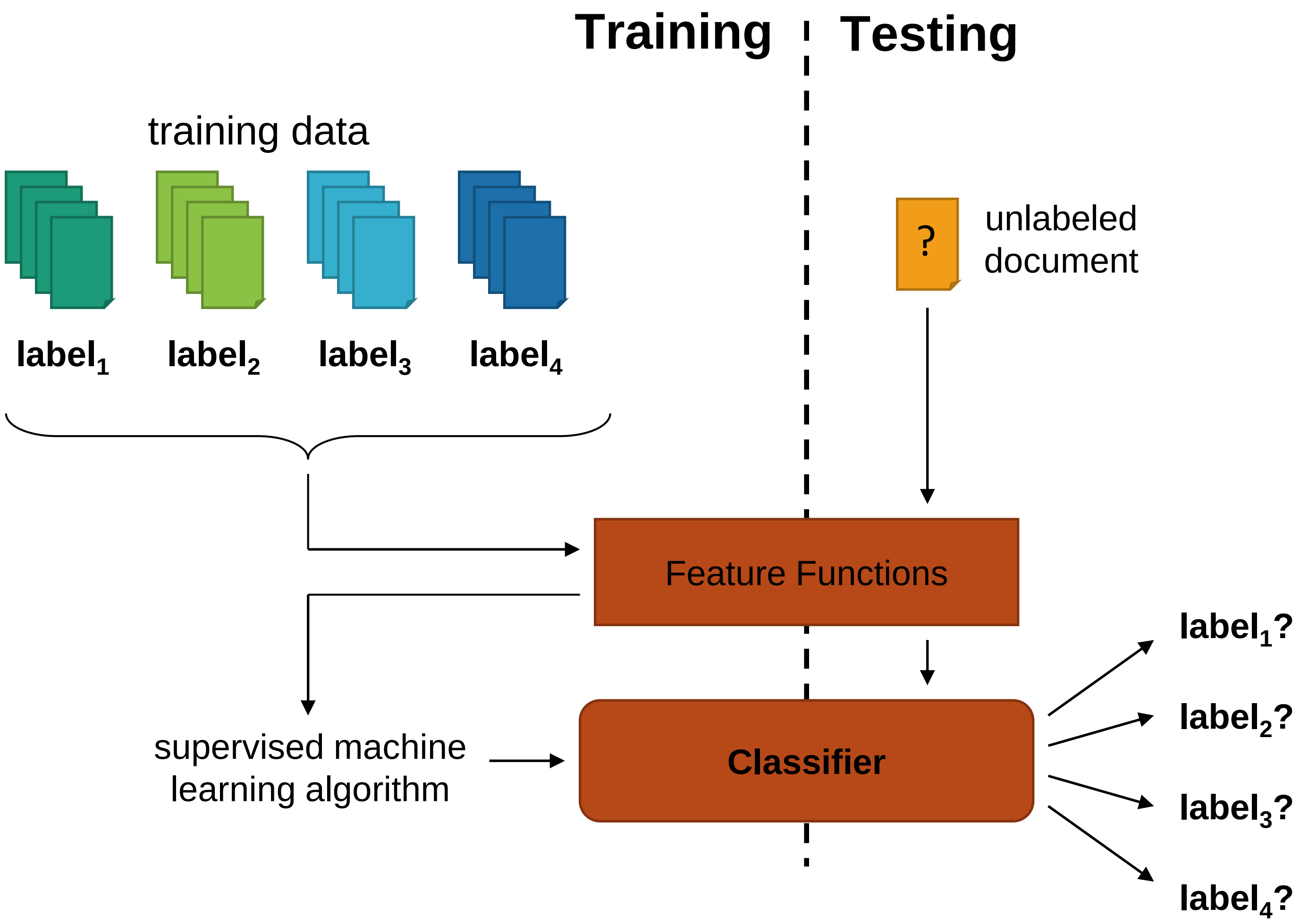
- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{a\frac{1}{P} + (1 - a)\frac{1}{R}} = \frac{(b^2 + 1)PR}{b^2 P + R}$$

- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$):

    $F = 2PR/(P+R)$

Multiclass Classification



Is this spam?

From: "Fabian Starr“ <Patrick_Freeman@pamietaniepeerelu.pl>

Subject: Hey! Sofware for the funny prices!

Get the great discounts on popular software today for PC and Macintosh
http://iiled.org/Cj4Lmx

70-90% Discounts from retail price!!!
All sofware is instantly available to download - No Need Wait!

What is the subject of this article?

MEDLINE Article


Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Word Sense Disambiguation what you should know

- Word senses distinguish different meanings of same word
- Sense inventories
- Annotation issues and annotator agreement (Kappa)
- Definition of Word Sense Disambiguation Task
- An unsupervised approach: Lesk algorithm
- Supervised classification:
  - Train vs. test data
  - The most frequent class baseline
- Evaluation metrics: accuracy, precision, recall