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)

- Homophones: same pronunciation, different orthography, different meaning
  - Examples: would/wood, to/too/two

- Homographs: distinct senses, same orthographic form, different pronunciation
  - Examples: bass (fish) vs. bass (instrument)

Relationship Between Senses

- **IS-A relationships**
  - From specific to general (up): hypernym (hypernymy)
  - From general to specific (down): hyponym (hyponymy)
- **Part-Whole relationships**
  - wheel is a meronym of car (meronymy)
  - car is a holonym of wheel (holonymy)

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

Word Sense Disambiguation

- Task: 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 big 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

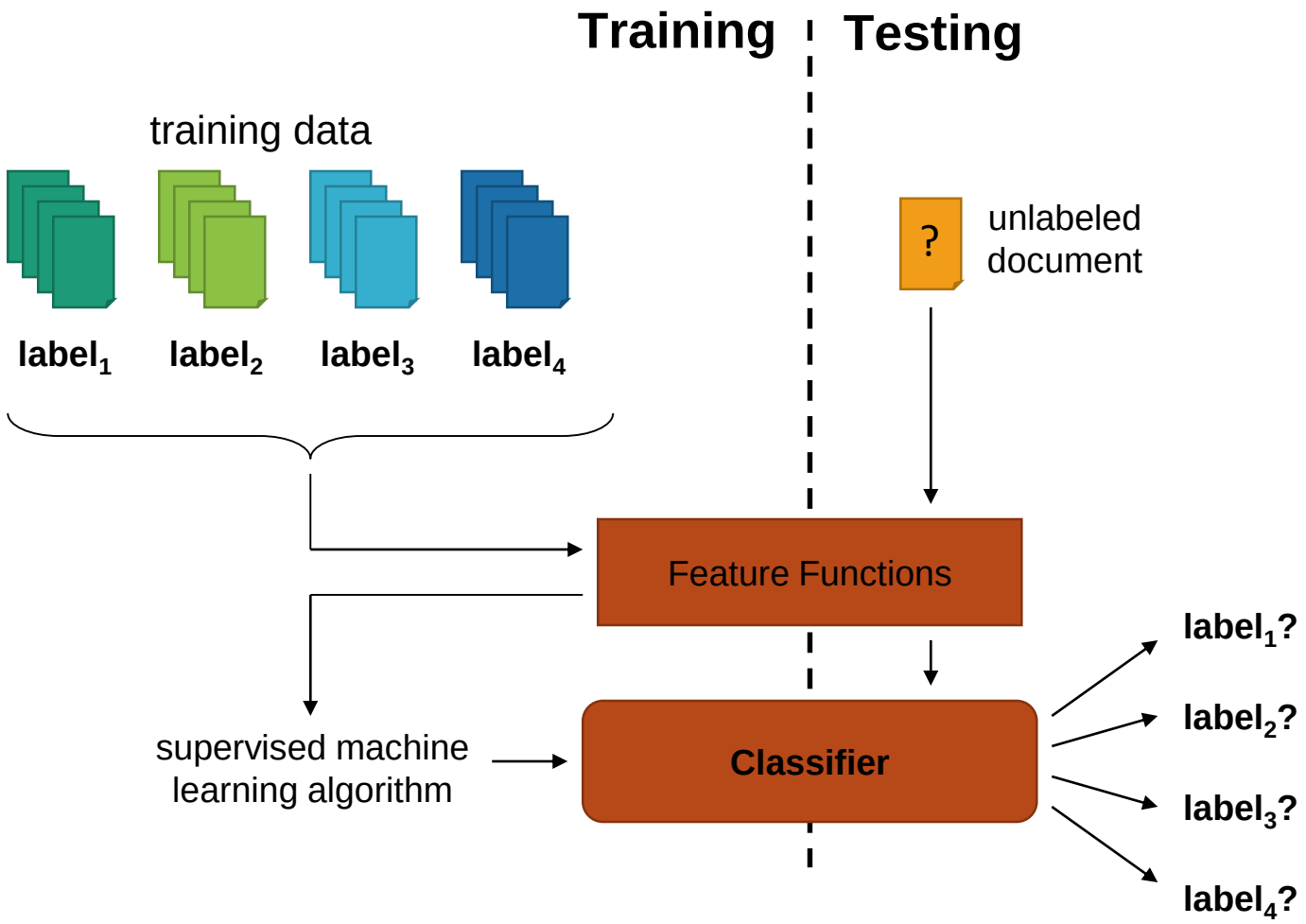
The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

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 (e.g., bigrams)
  - Give extra weight to infrequent words
  - …

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)
  - …

Word Meaning

2 core issues from an NLP perspective

- **Semantic similarity**: given two words, how similar are they in meaning?
- Key concepts: vector semantics, PPMI and its variants, cosine similarity

- **Word sense disambiguation**: given a word that has more than one meaning, which one is used in a specific context?
- Key concepts: word sense, WordNet and sense inventories, unsupervised disambiguation (Lesk), supervised disambiguation