TTIC 31190: Natural Language Processing

Kevin Gimpel

Winter 2016

Lecture 4: Lexical Semantics

Roadmap

- classification
- words
- lexical semantics
- language modeling
- sequence labeling
- syntax and syntactic parsing
- neural network methods in NLP
- semantic compositionality
- semantic parsing
- unsupervised learning
- machine translation and other applications

Why is NLP hard?

- ambiguity and variability of linguistic expression:
  - **ambiguity**: one form can mean many things
  - **variability**: many forms can mean the same thing

Feature Engineering for Text Classification

- Two features:

$$f_1(x, y) = \mathbb{I}[y = \text{ positive}] \wedge \mathbb{I}[x \text{ contains } \mathit{great}]$$
$$f_2(x, y) = \mathbb{I}[y = \text{ negative}] \wedge \mathbb{I}[x \text{ contains } \mathit{great}]$$

ambiguity: "***great***" can mean different things in different contexts

- On sentences containing "***great***" in the Stanford Sentiment Treebank training data, this would get us an accuracy of 69%
- But "***great***'' only appears in 83/6911 examples

variability: many other words can indicate positive sentiment

- most of what we talked about on Tuesday and what we will talk about today:

- one form, multiple meanings → split form

ambiguity

- multiple forms, one meaning → merge forms

variability

Ambiguity

- one form, multiple meanings → split form
  - tokenization (adding spaces):
    - *didn’t → did n’t*
    - *“Yes?” → “ Yes ? ”*
  - **today**: word sense disambiguation:
    - *power plant → power plant$_1$*
    - *flowering plant → flowering plant$_2$*

Variability

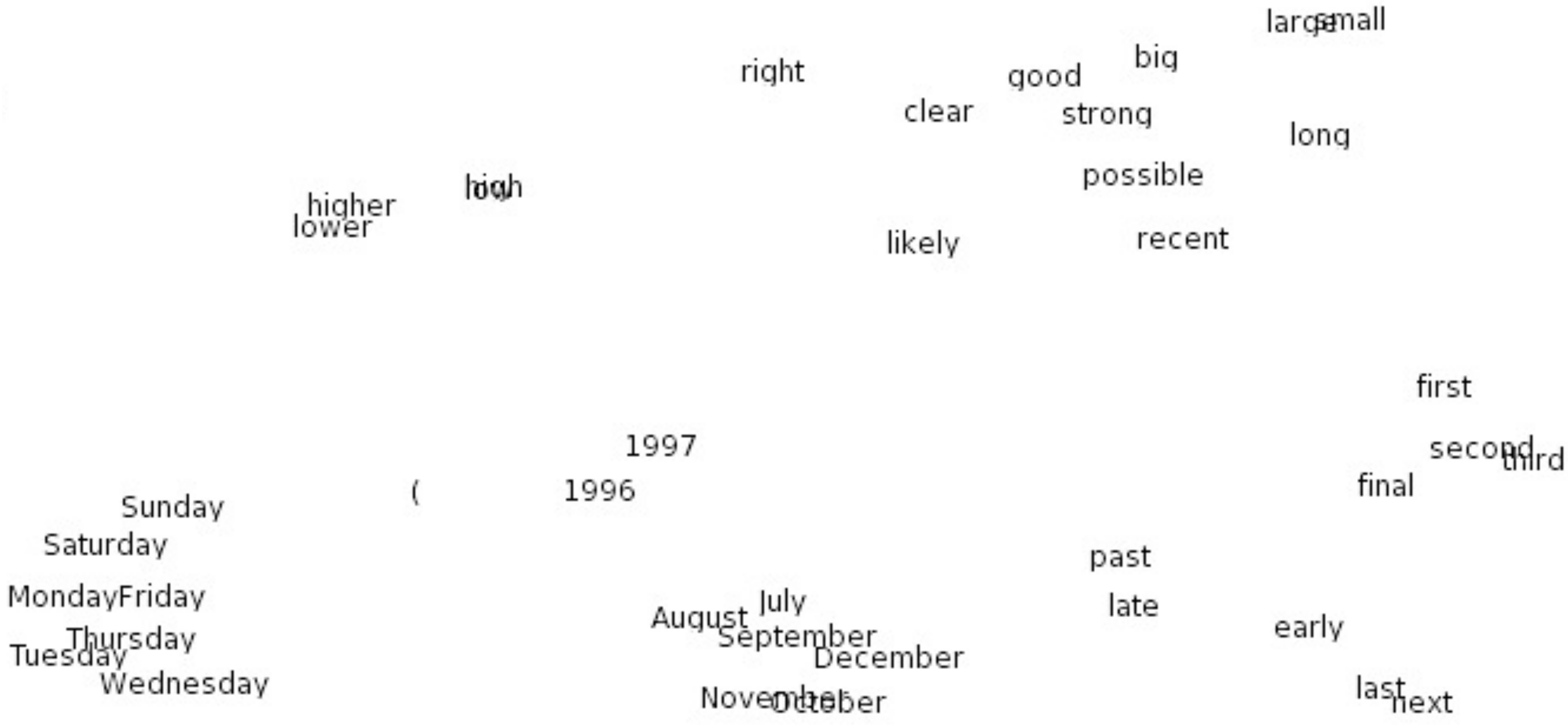
- multiple forms, one meaning → merge forms
  - tokenization (removing spaces):
    - *New York → NewYork*
  - lemmatization:
    - *walked → walk*
    - *walking → walk*
  - stemming:
    - *automation → automat*
    - *automates → automat*

are there any NLP tasks where doing this might lose valuable information?

Variability

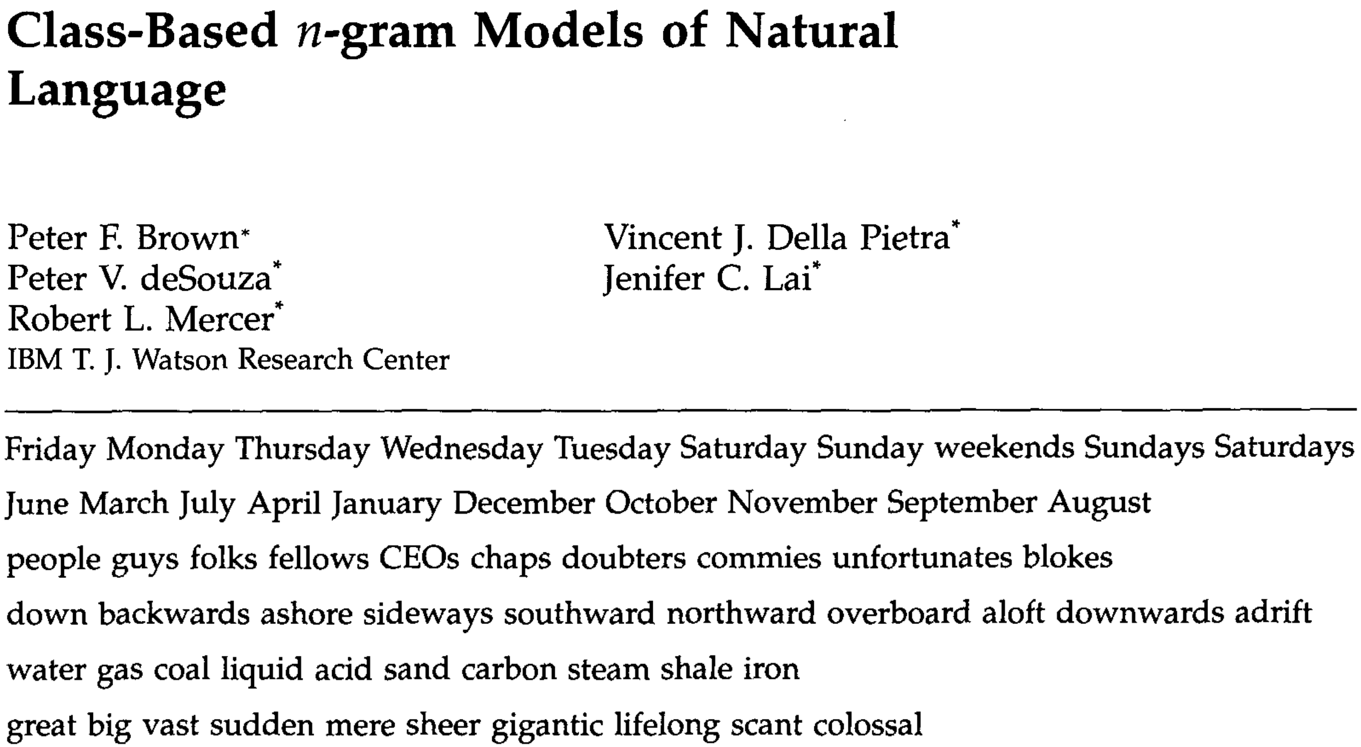
- multiple forms, one meaning → merge forms
  - tokenization (removing spaces):
    - *New York → NewYork*
  - lemmatization:
    - *walked → walk*
    - *walking → walk*
  - stemming:
    - *automation → automat*
    - *automates → automat*
  - **today/next week**: word representations

Vector Representations of Words



t-SNE visualization from Turian et al. (2010)

Word Clusters



Class-Based *n*-gram Models of Natural Language

Peter F. Brown*
Peter V. deSouza*
Robert L. Mercer*
IBM T. J. Watson Research Center

Vincent J. Della Pietra*
Jenifer C. Lai*

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays

June March July April January December October November September August

people guys folks fellows CEOs chaps doubters commies unfortunates blokes

down backwards ashore sideways southward northward overboard aloft downwards adrift

water gas coal liquid acid sand carbon steam shale iron

great big vast sudden mere sheer gigantic lifelong scant colossal

Computational Linguistics, 1992

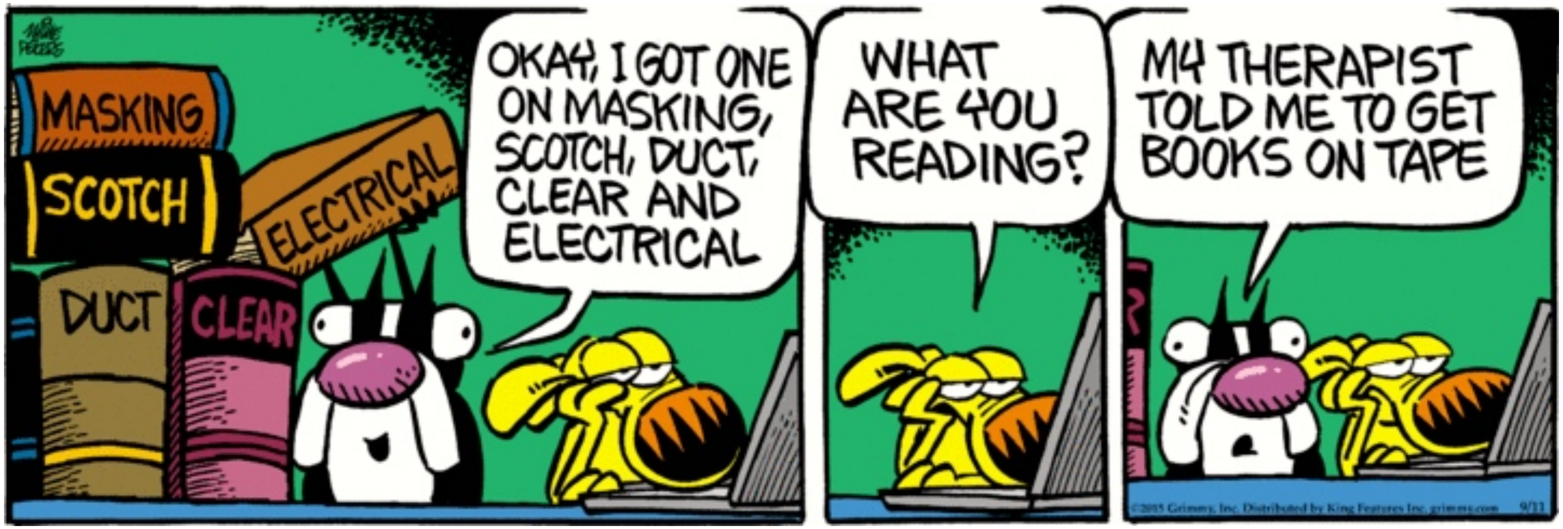
Roadmap

- lexical semantics
  - word sense
  - word sense disambiguation
  - word representations

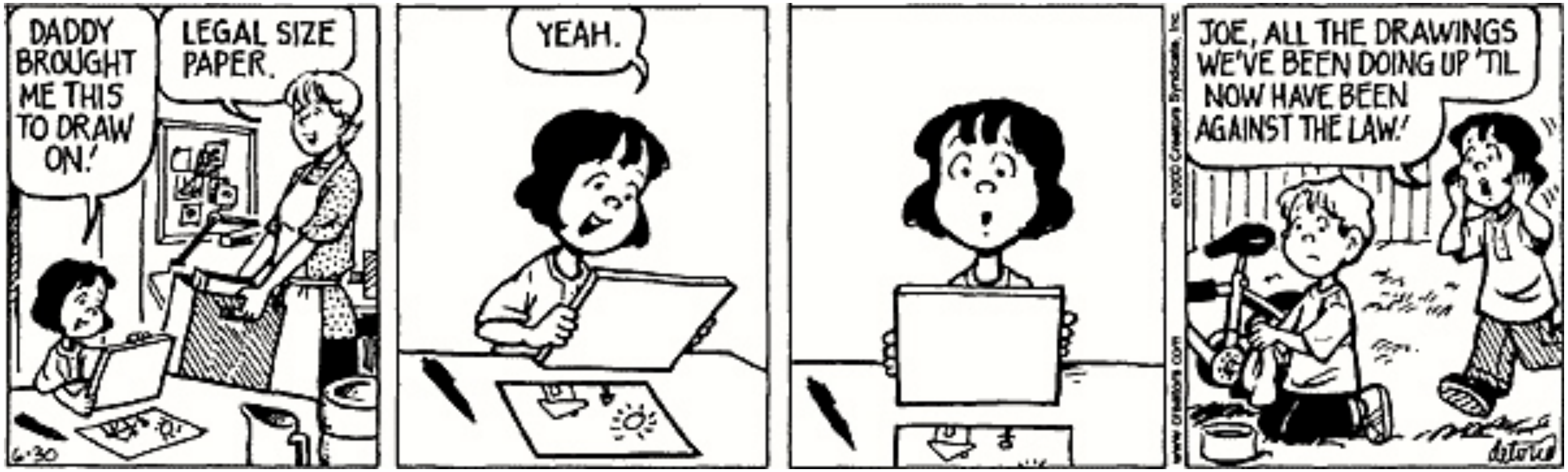
Word Sense Ambiguity

- many words have multiple meanings

Word Sense Ambiguity


MASKING
SCOTCH
DUCT
CLEAR
ELECTRICAL
OKAY, I GOT ONE ON MASKING, SCOTCH, DUCT, CLEAR AND ELECTRICAL
WHAT ARE YOU READING?
MY THERAPIST TOLD ME TO GET BOOKS ON TAPE
©2015 Grimmy, Inc. Distributed by King Features Inc. grimmy.com 9/11

credit: A. Zwicky


DADDY BROUGHT ME THIS TO DRAW ON!
LEGAL SIZE PAPER.
YEAH.
JOE, ALL THE DRAWINGS WE'VE BEEN DOING UP 'TIL NOW HAVE BEEN AGAINST THE LAW!
6-30
©2000 Creators Syndicate, Inc
www.creators.com
detorie

Terminology: lemma and wordform

- **lemma**
  - words with same lemma have same stem, part of speech, rough semantics
- **wordform**
  - inflected word as it appears in text

wordform	lemma
banks	bank
sung	sing
duermes	dormir

J&M/SLP3

Lemmas have senses

- one lemma *bank* can have many meanings:

sense 1: …a *bank*$_1$ can hold the investments in a custodial account

sense 2: …as agriculture burgeons on the east *bank*$_2$ the river will shrink even more

- **sense** (or **word sense**)
  - a discrete representation of an aspect of a word's meaning
- the lemma *bank* here has two senses

- two ways to categorize the patterns of multiple meanings of words:
  - **homonymy**: the multiple meanings are unrelated (coincidental?)
  - **polysemy**: the multiple meanings are related

Homonymy

homonyms: words that share a form but have unrelated, distinct meanings:

- *bank*$_1$: financial institution, *bank*$_2$: sloping land
- *bat*$_1$: club for hitting a ball, *bat*$_2$: nocturnal flying mammal

homographs: same spelling, different meanings

bank/bank, bat/bat

homophones: same pronunciation, different meanings

write/right, piece/peace

J&M/SLP3

Homonymy causes problems for NLP

- information retrieval
  - query: *bat care*
- machine translation
  - *bat*: *murciélago* (animal) or *bate* (for baseball)
- text-to-speech
  - *bass* (stringed instrument) vs. *bass* (fish)

Polysemy

1: *The bank was constructed in 1875 out of local red brick.*

2: *I withdrew the money from the bank*.

- are these the same sense?
  - sense 2: “a financial institution”
  - sense 1: “the building belonging to a financial institution”

- a **polysemous** word has **related** meanings
  - most non-rare words have multiple meanings

J&M/SLP3

Metonymy or Systematic Polysemy: A systematic relationship between senses

- lots of types of polysemy are systematic
  - *school, university, hospital*
  - all can mean the institution or the building
- a systematic relationship:
  - *building* ⟷ *organization*
- other such kinds of systematic polysemy:

Author (Jane Austen wrote Emma) ⟷ *Works of Author* (I love Jane Austen)

Tree (Plums have beautiful blossoms) ⟷ *Fruit* (I ate a preserved plum)

J&M/SLP3

How do we know when a word has more than one sense?

- "zeugma" test: two senses of `serve`?
  - *Which flights* ***serve*** *breakfast?*
  - *Does Lufthansa* ***serve*** *Philadelphia?*
  - *?Does Lufthansa serve breakfast and Philadelphia?*
- since this conjunction sounds weird, we say that these are **two different senses of *serve***

Synonyms

- words with same meaning in some or all contexts:
  - *filbert / hazelnut*
  - *couch / sofa*
  - *big / large*
  - *automobile / car*
  - *vomit / throw up*
  - *water /* H_20
- two lexemes are synonyms if they can be substituted for each other in all situations

J&M/SLP3

Synonyms

- few (or no) examples of perfect synonymy
  - even if many aspects of meaning are identical
  - still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- example:
  - *water / H_20*
  - *big / large*
  - *brave / courageous*

J&M/SLP3

Synonymy is a relation between **senses** rather than words

- consider the words *big* and *large*
- are they synonyms?
  - How ***big*** is that plane?
  - Would I be flying on a ***large*** or small plane?
- how about here:
  - Miss Nelson became a kind of ***big*** sister to Benjamin.
  - ?Miss Nelson became a kind of ***large*** sister to Benjamin.
- why?
  - *big* has a sense that means being older or grown up
  - *large* lacks this sense

J&M/SLP3

Antonyms

- senses that are opposites with respect to one feature of meaning
- otherwise, they are very similar!

  dark/light   short/long   fast/slow   rise/fall

  hot/cold   up/down   in/out

- more formally, antonyms can
  - define a binary opposition or be at opposite ends of a scale
    - *long/short, fast/slow*
  - be **reversives**:
    - *rise/fall, up/down*

Hyponymy and Hypernymy

- one sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *mango* is a hyponym of *fruit*
- conversely: **hypernym** ("hyper is super")
  - *vehicle* is a hypernym of *car*
  - *fruit* is a hypernym of *mango*

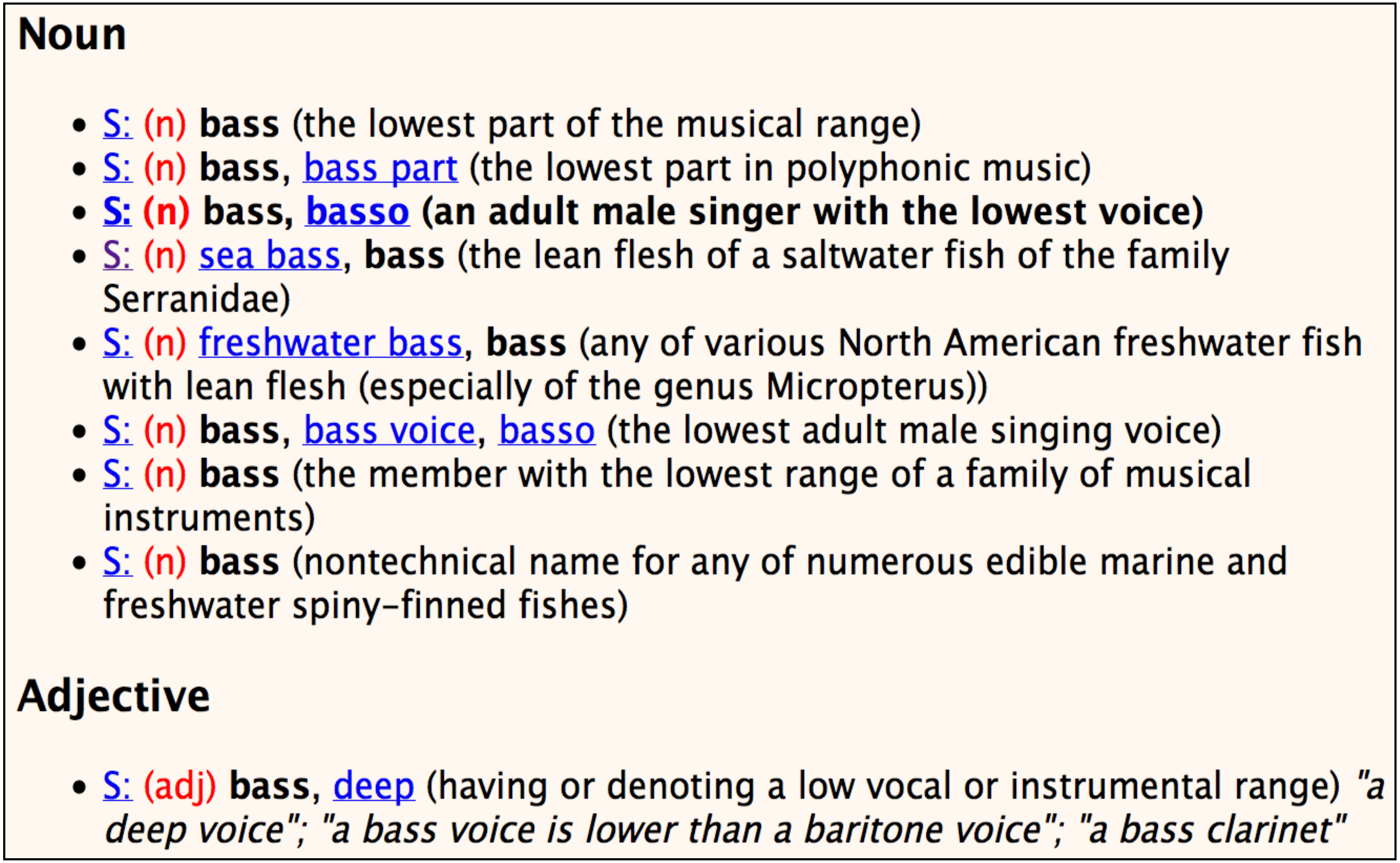
WordNet 3.0

- hierarchically organized lexical database
- on-line thesaurus + aspects of a dictionary
  - some languages available or under development: Arabic, Finnish, German, Portuguese…

Category	Unique Strings
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

J&M/SLP3

Senses of *bass* in WordNet



Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- **S: (n) bass, basso (an adult male singer with the lowest voice)**
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

J&M/SLP3

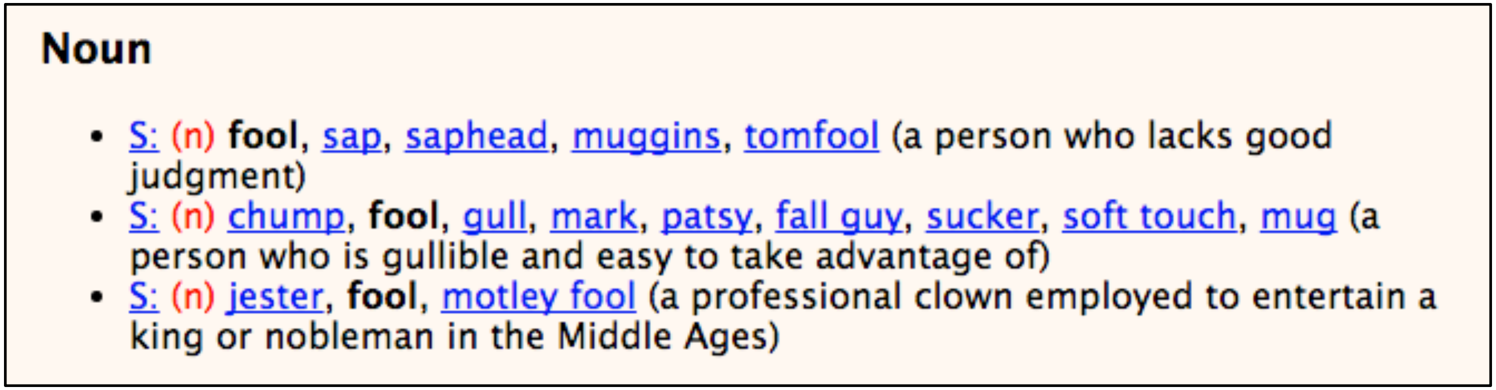
How is "sense" defined in WordNet?

- **synset (synonym set)**: set of near-synonyms; instantiates a sense or concept, with a **gloss**
- example: *chump* as a noun with gloss:

  "a person who is gullible and easy to take advantage of"
- this sense of *chump* is shared by 9 words:

  chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug^2
- each of **these** senses have this same gloss
  - (not **every** sense; sense 2 of *gull* is the aquatic bird)

J&M/SLP3



Noun

- S: (n) **fool**, sap, saphead, muggins, tomfool (a person who lacks good judgment)
- S: (n) chump, **fool**, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)
- S: (n) jester, **fool**, motley fool (a professional clown employed to entertain a king or nobleman in the Middle Ages)

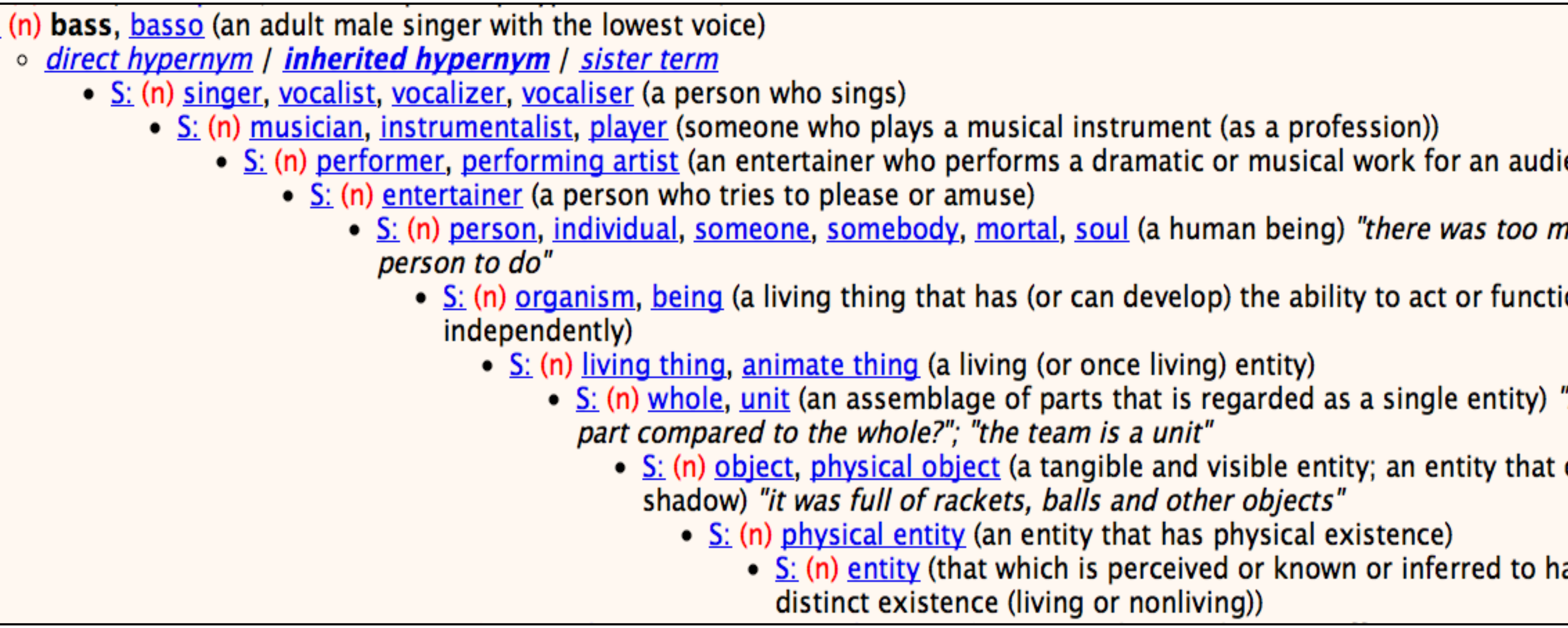
ambiguity

- one form, multiple meanings → split form
  - the three senses of *fool* belong to different synsets

variability

- multiple forms, one meaning → merge forms
  - each synset contains senses of several different words

WordNet Hypernym Hierarchy for *bass*



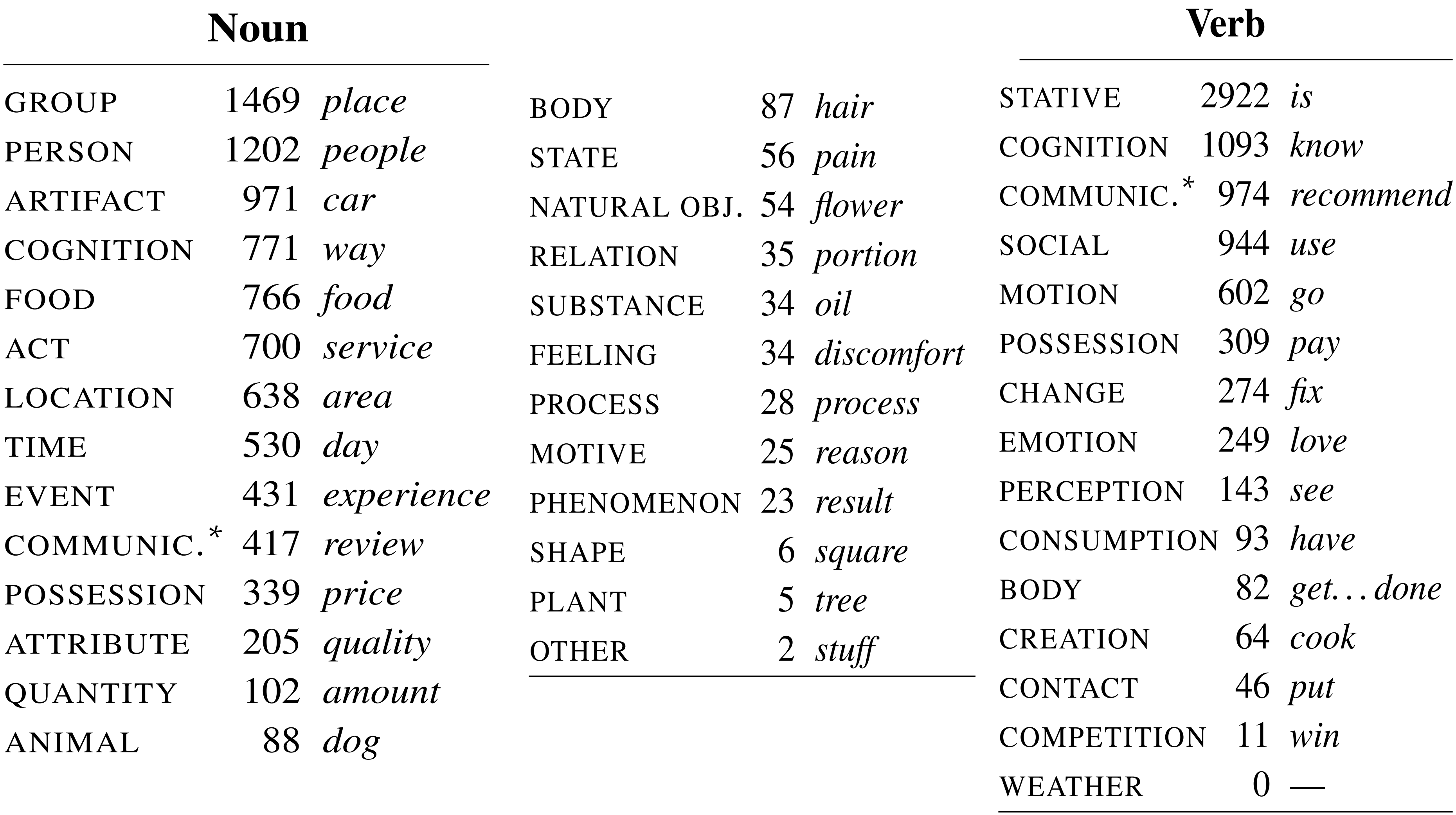
(n) **bass**, basso (an adult male singer with the lowest voice)

- *direct hypernym* / ***inherited hypernym*** / *sister term*
  - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
    - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
      - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audi
        - S: (n) entertainer (a person who tries to please or amuse)
          - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too m person to do"*
            - S: (n) organism, being (a living thing that has (or can develop) the ability to act or functi independently)
              - S: (n) living thing, animate thing (a living (or once living) entity)
                - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *" part compared to the whole?"; "the team is a unit"*
                  - S: (n) object, physical object (a tangible and visible entity; an entity that c shadow) *"it was full of rackets, balls and other objects"*
                    - S: (n) physical entity (an entity that has physical existence)
                      - S: (n) entity (that which is perceived or known or inferred to ha distinct existence (living or nonliving))

J&M/SLP3

Supersenses: top level hypernyms in hierarchy

(counts from Schneider & Smith's Streusel corpus)



Noun

GROUP	1469	*place*
PERSON	1202	*people*
ARTIFACT	971	*car*
COGNITION	771	*way*
FOOD	766	*food*
ACT	700	*service*
LOCATION	638	*area*
TIME	530	*day*
EVENT	431	*experience*
COMMUNIC.*	417	*review*
POSSESSION	339	*price*
ATTRIBUTE	205	*quality*
QUANTITY	102	*amount*
ANIMAL	88	*dog*
BODY	87	*hair*
STATE	56	*pain*
NATURAL OBJ.	54	*flower*
RELATION	35	*portion*
SUBSTANCE	34	*oil*
FEELING	34	*discomfort*
PROCESS	28	*process*
MOTIVE	25	*reason*
PHENOMENON	23	*result*
SHAPE	6	*square*
PLANT	5	*tree*
OTHER	2	*stuff*

Verb

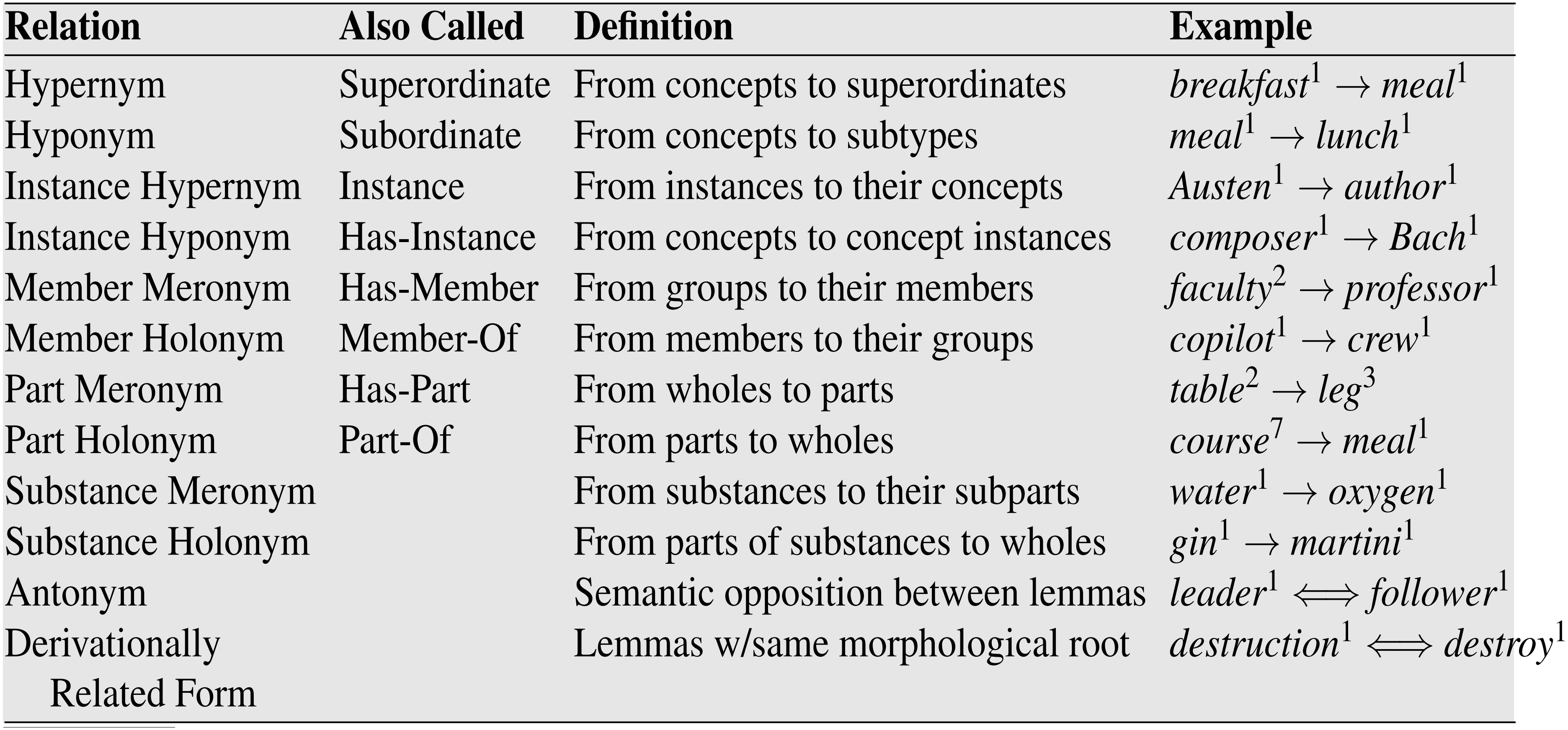
STATIVE	2922	*is*
COGNITION	1093	*know*
COMMUNIC.*	974	*recommend*
SOCIAL	944	*use*
MOTION	602	*go*
POSSESSION	309	*pay*
CHANGE	274	*fix*
EMOTION	249	*love*
PERCEPTION	143	*see*
CONSUMPTION	93	*have*
BODY	82	*get...done*
CREATION	64	*cook*
CONTACT	46	*put*
COMPETITION	11	*win*
WEATHER	0	—

Meronymy

- part-whole relation
  - *wheel* is a **meronym** of *car*
  - *car* is a **holonym** of *wheel*

J&M/SLP3

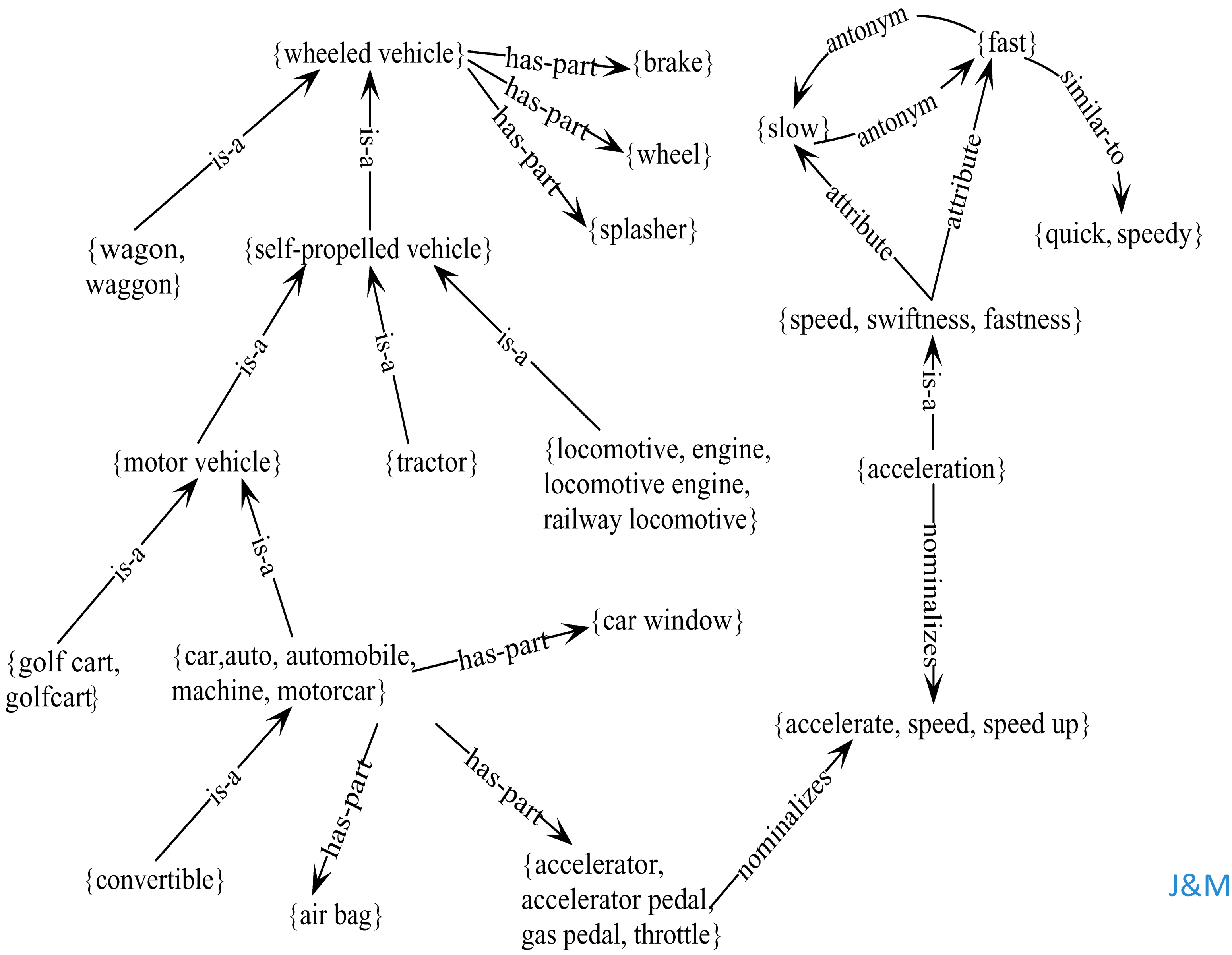
WordNet Noun Relations



Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} → *author*^{1}
Instance Hyponym	Has-Instance	From concepts to concept instances	*composer*^{1} → *Bach*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Substance Meronym		From substances to their subparts	*water*^{1} → *oxygen*^{1}
Substance Holonym		From parts of substances to wholes	*gin*^{1} → *martini*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} ⟺ *follower*^{1}
Derivationally Related Form		Lemmas w/same morphological root	*destruction*^{1} ⟺ *destroy*^{1}

J&M/SLP3

WordNet: Viewed as a graph

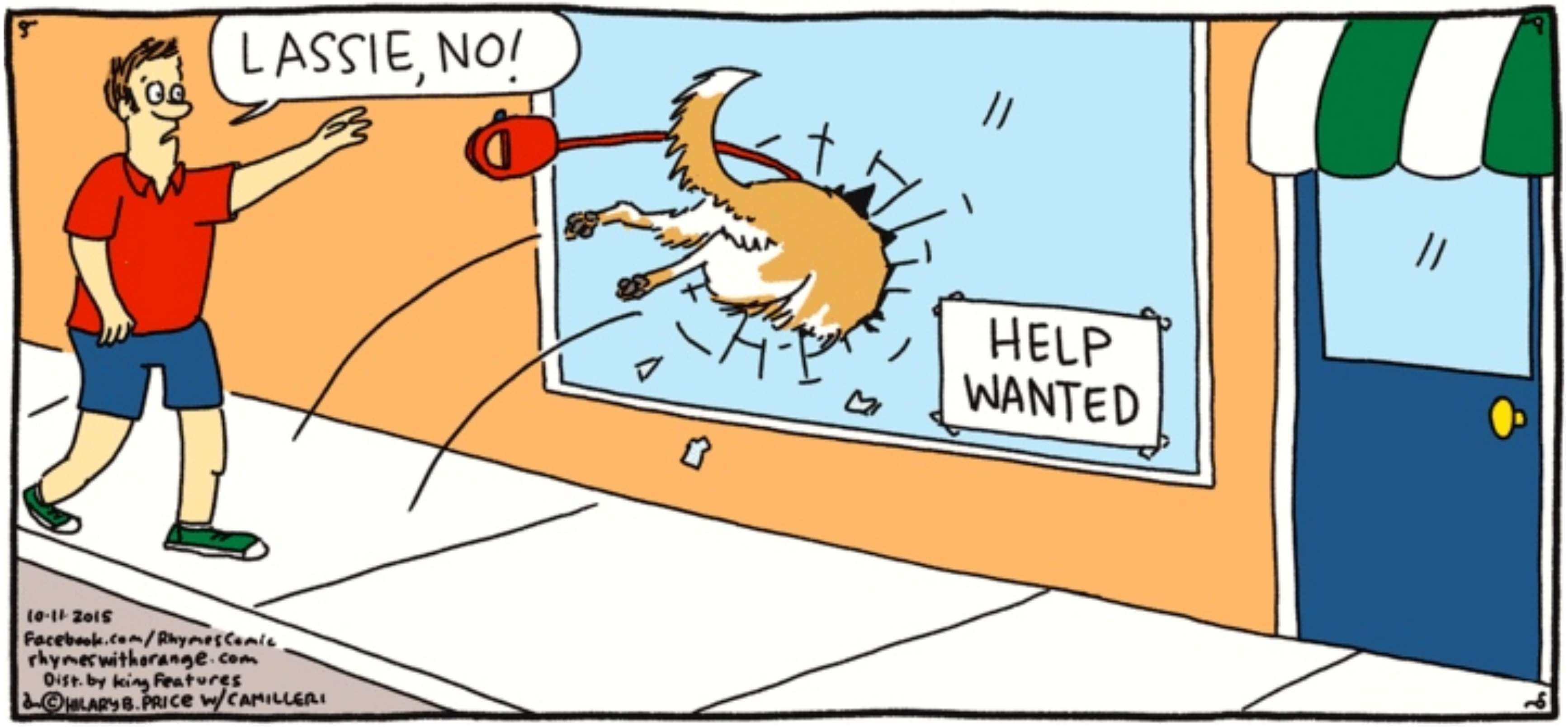


J&M/SLP3

Roadmap

- lexical semantics
  - word sense
  - word sense disambiguation
  - word representations

Word Sense Disambiguation



Word Sense Disambiguation (WSD)

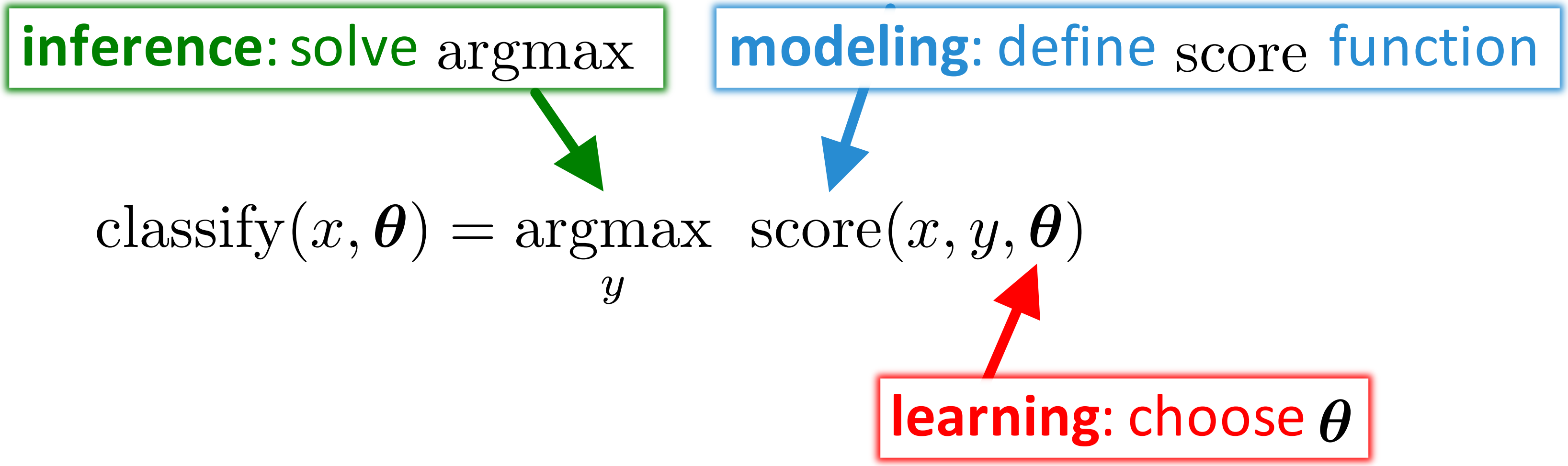
- given:
  - a word in context
  - a fixed inventory of potential word senses
- decide which sense of the word this is
- why? machine translation, question answering, sentiment analysis, text-to-speech
- what set of senses?
  - English-to-Spanish machine translation: set of Spanish translations
  - text-to-speech: homographs like *bass* and *bow*
  - in general: the senses in a thesaurus like WordNet

J&M/SLP3

Two Variants of WSD Task

- lexical sample task
  - small pre-selected set of target words (*line, plant, bass*)
  - inventory of senses for each word
  - supervised learning: train a classifier for **each** word
- all-words task
  - every word in an entire text
  - a lexicon with senses for each word
  - data sparseness: can't train word-specific classifiers

Classification Framework



inference: solve argmax

modeling: define score function

$$\mathrm{classify}(x, \boldsymbol{\theta}) = \operatorname*{argmax}_{y} \; \mathrm{score}(x, y, \boldsymbol{\theta})$$

learning: choose $\boldsymbol{\theta}$

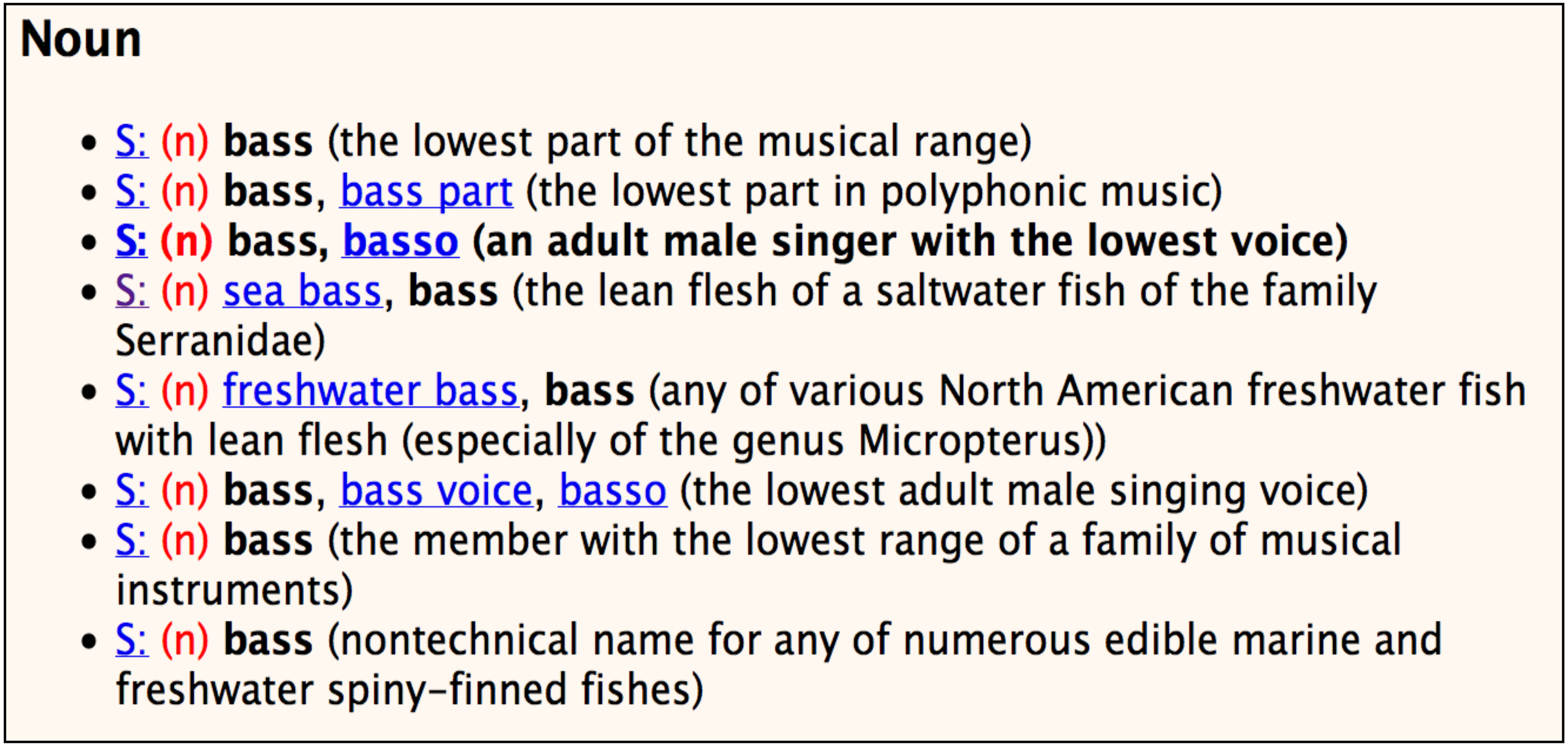
Classification for Word Sense Disambiguation of *bass*

$$\text{classify}(x, \boldsymbol{\theta}) = \operatorname*{argmax}_{y} \ \text{score}(x, y, \boldsymbol{\theta})$$

- **data**:
  - what is the space of possible inputs and outputs?
  - what do (*x*,*y*) pairs look like?
  - *x* = the word *bass* along with its context
  - *y* = word sense of *bass* (from a list of possible senses)

$$\text{classify}_{\text{bassWSD}}^{\text{linear}}(\boldsymbol{x}, \boldsymbol{\theta}) = \operatorname*{argmax}_{y \in \mathcal{L}_{\text{bass}}} \ \sum_{i} \theta_i f_i(\boldsymbol{x}, y)$$

8 Senses of *bass* in Wordnet

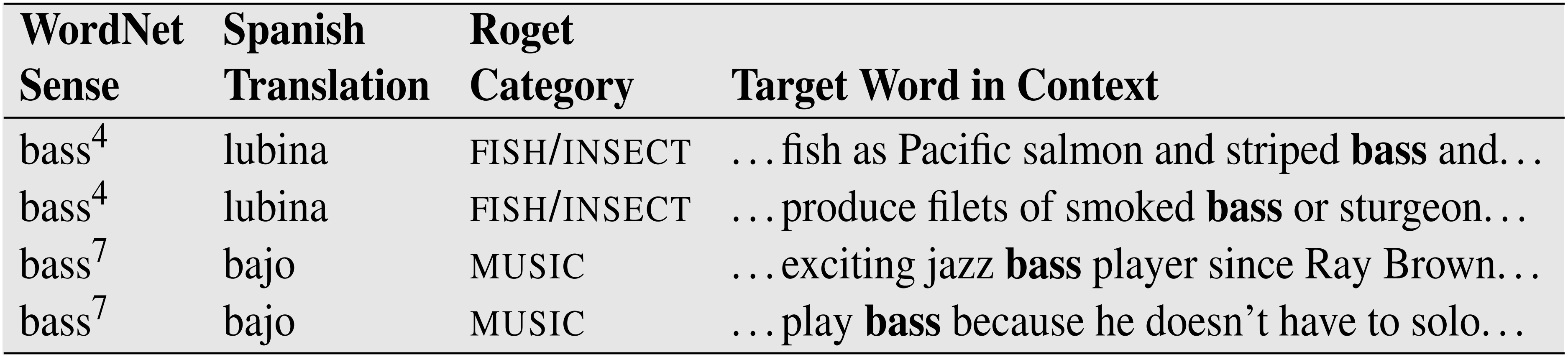


Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- **S: (n) bass, basso (an adult male singer with the lowest voice)**
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

J&M/SLP3

Inventory of Sense Tags for *bass*



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…

J&M/SLP3

WSD Evaluation and Baselines

- best evaluation: **extrinsic ("task-based")**
  - embed WSD in a task and see if it helps!

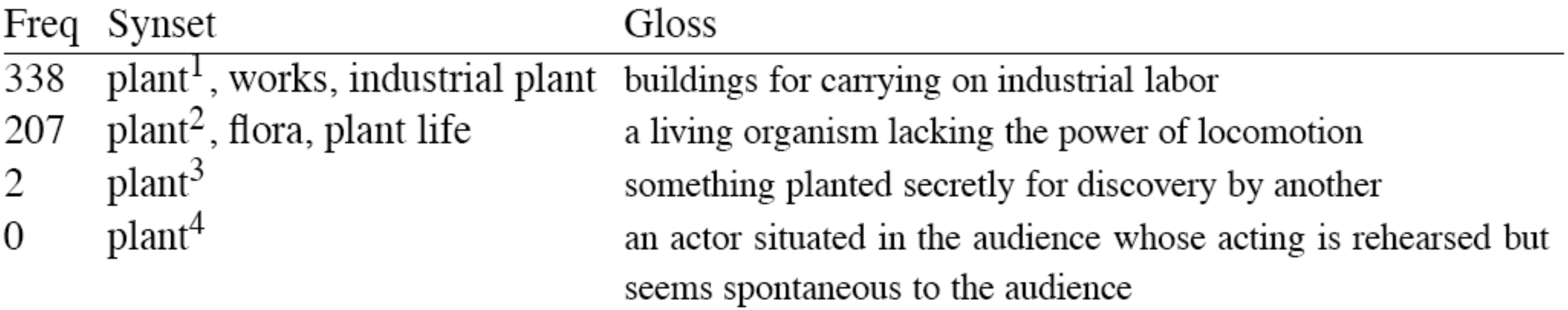
- **intrinsic** evaluation often done for convenience

- strong baseline: most frequent sense

J&M/SLP3

Most Frequent Sense

- WordNet senses are ordered by frequency
- most frequent is first
- sense frequencies come from *SemCor* corpus



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

J&M/SLP3

Performance Ceiling

- human inter-annotator agreement
  - compare annotations of two humans on same data, given same tagging guidelines

- human agreements on all-words corpora with WordNet style senses: 75%-80%

J&M/SLP3

Training Data for WSD

- **semantic concordance**: corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus
  - SemCor: 234,000 words from Brown Corpus, manually tagged with WordNet senses
  - SENSEVAL-3 competition corpora: 2081 tagged word tokens

Features for WSD?

Features for WSD?
Intuition from Warren Weaver (1955):

“If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…

The practical question is : ‘What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?’”

J&M/SLP3

Features for WSD

- **collocational**
  - features for words at **specific** positions near target word
- **bag-of-words**
  - features for words that occur anywhere in a window of the target word (regardless of position)

J&M/SLP3

Example

- using a window of +/- 3 from the target:

  An electric guitar and ***bass*** *player stand off* to *one side not really part of the scene*

Semi-Supervised Learning

problem: supervised learning requires large hand-built resources

what if you don’t have much training data?

solution: bootstrapping

generalize from a very small hand-labeled seed set

J&M/SLP3

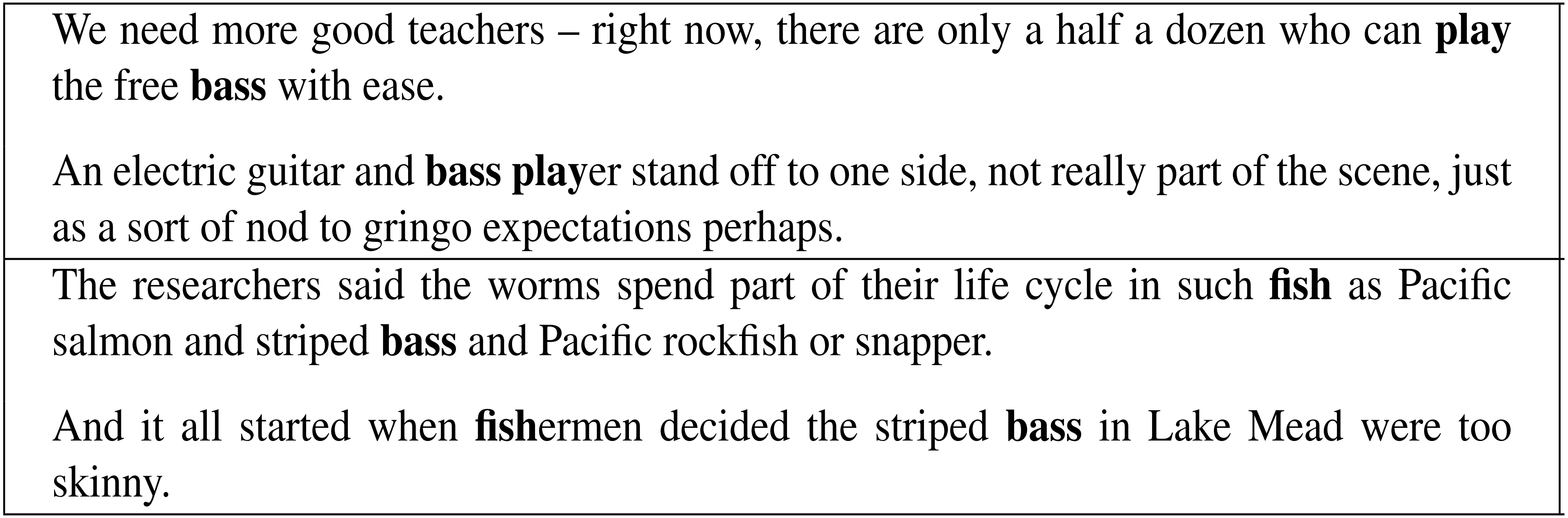
Bootstrapping

- “one sense per collocation” heuristic:
  - – a word reoccurring in collocation with the same word will almost surely have the same sense

- For *bass*:
  - – *play* occurs with the music sense of *bass*
  - – *fish* occurs with the fish sense of *bass*

J&M/SLP3

Sentences extracted using *fish* and *play*



We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease.

An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper.

And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

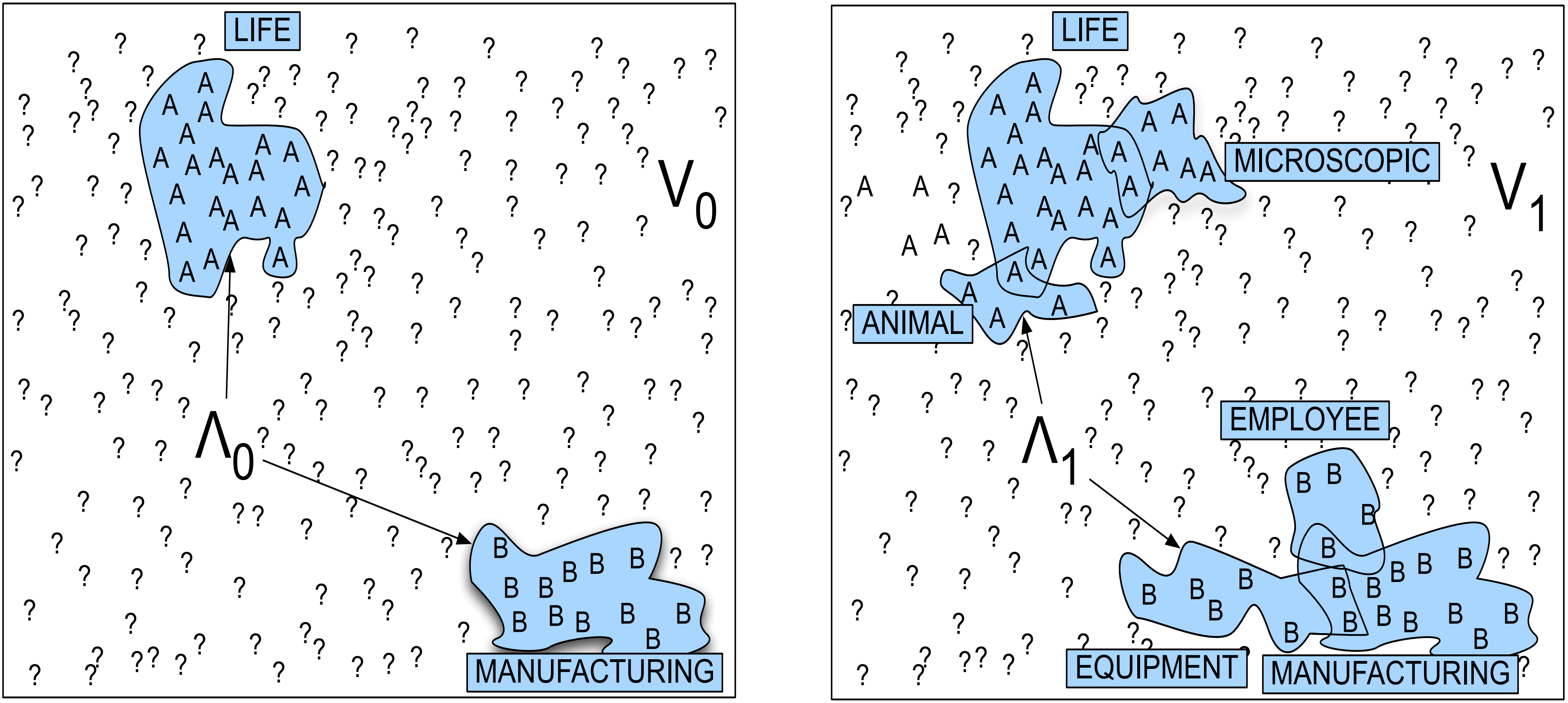
J&M/SLP3

Bootstrapping

- “one sense per collocation” heuristic:
  - a word reoccurring in collocation with the same word will almost surely have the same sense

- “one sense per discourse” heuristic:
  - sense of a word is highly consistent within a document (Yarowsky, 1995)
  - especially topic-specific words

Stages in Yarowsky bootstrapping algorithm for *plant*



(a) (b)

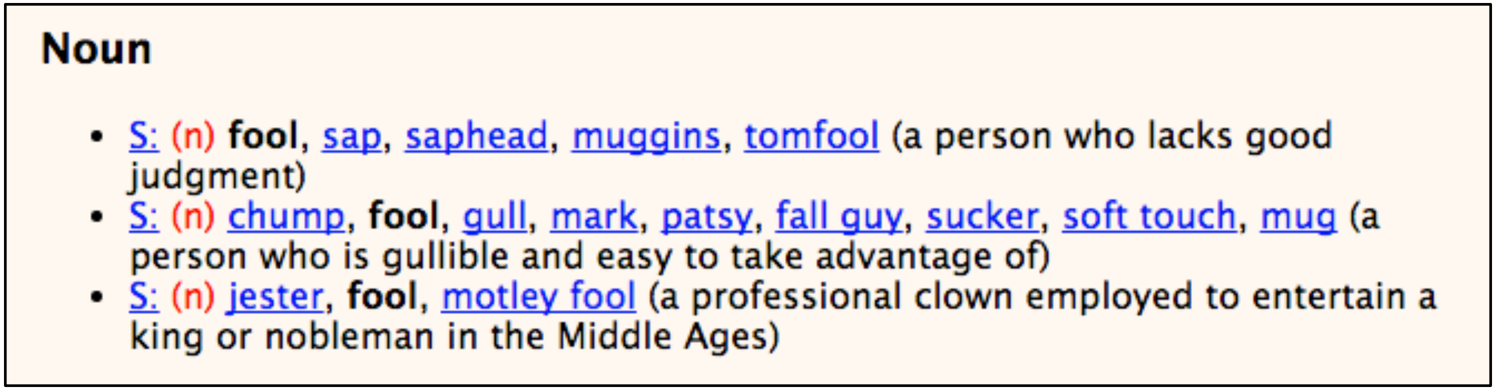
J&M/SLP3

Summary

- word sense disambiguation: choosing correct sense in context
- applications: MT, QA, etc.
- main intuition:
  - lots of information in a word’s context
  - simple algorithms based on word counts can be surprisingly good

Roadmap

- lexical semantics
  - word sense
  - word sense disambiguation
  - word representations



Noun

- S: (n) **fool**, sap, saphead, muggins, tomfool (a person who lacks good judgment)
- S: (n) chump, **fool**, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)
- S: (n) jester, **fool**, motley fool (a professional clown employed to entertain a king or nobleman in the Middle Ages)

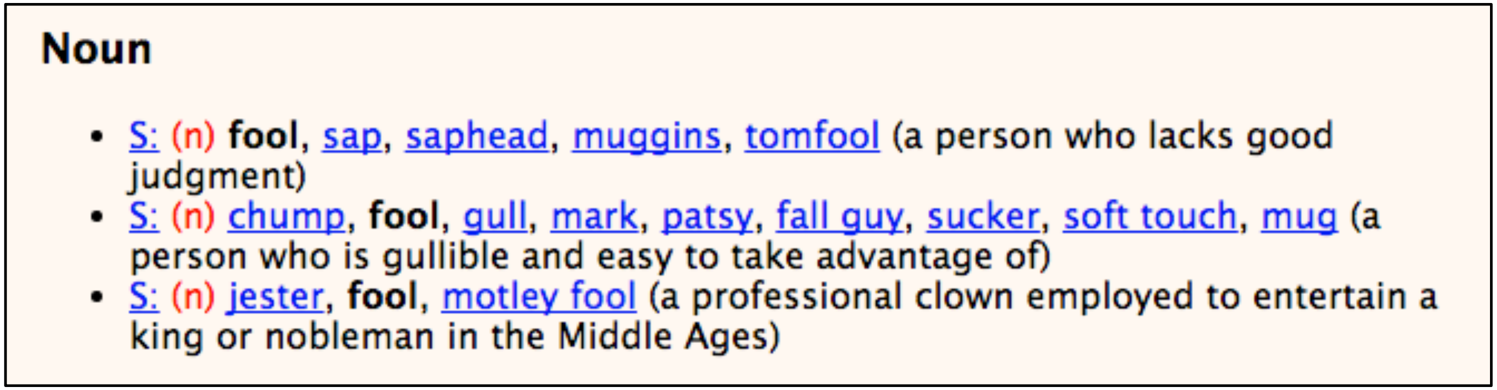
ambiguity

- one form, multiple meanings → split form
  - – the three senses of *fool* belong to different synsets

variability

- multiple forms, one meaning → merge forms
  - – each synset contains senses of several different words

- WordNet splits words into synsets; each contains senses of several words:



Noun

- S: (n) **fool**, sap, saphead, muggins, tomfool (a person who lacks good judgment)
- S: (n) chump, **fool**, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)
- S: (n) jester, **fool**, motley fool (a professional clown employed to entertain a king or nobleman in the Middle Ages)

- are we finished? have we solved the problem of representing word meaning?

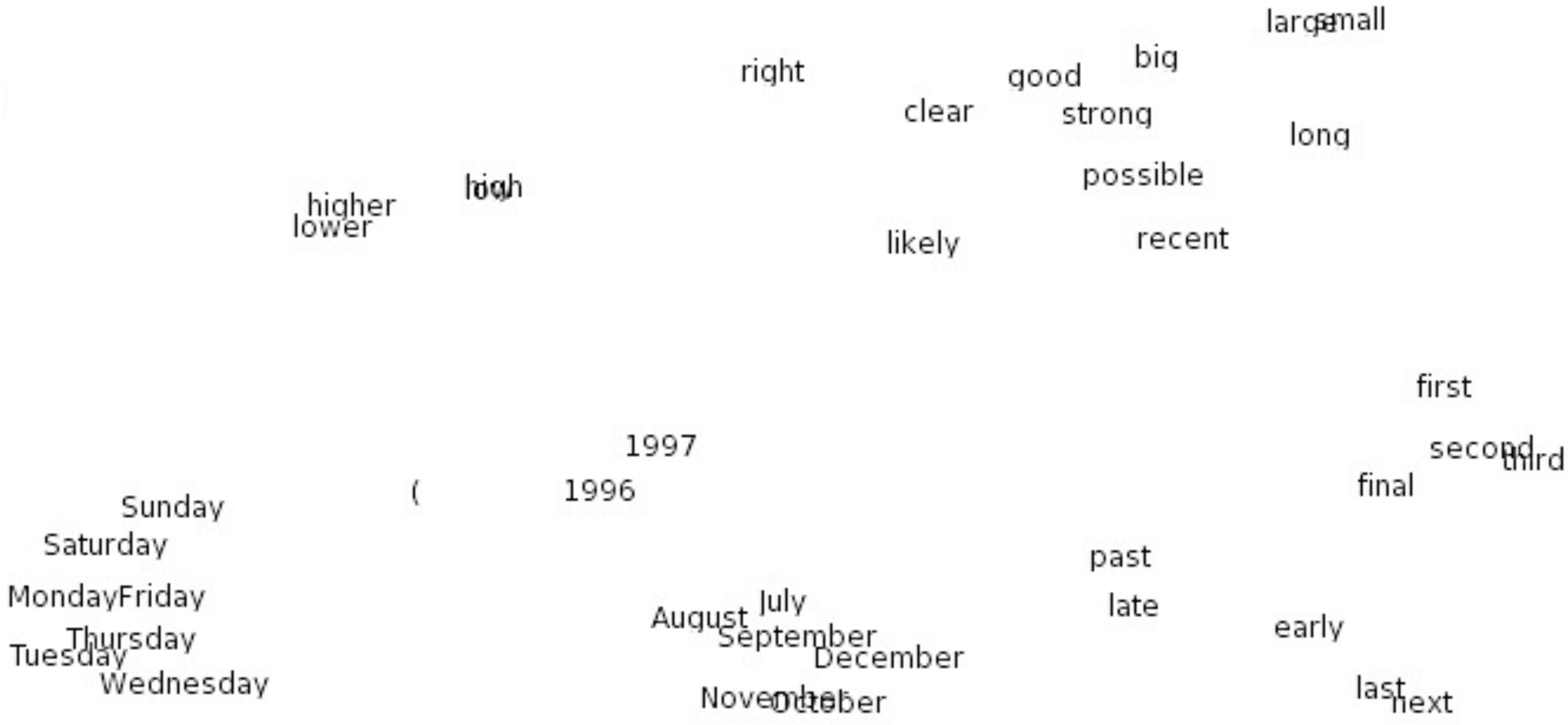
- are we finished? have we solved the problem of representing word meaning?
- issues:
  - WordNet has limited coverage and only exists for a small set of languages

category	unique strings
noun	117,798
verb	11,529
adjective	22,479
adverb	4,481

  - WSD requires training data, whether supervised or seeds for semi-supervised
- better approach: jointly learn representations for all words in an unsupervised way

Word Embeddings

(Bengio et al., 2003)



t-SNE visualization from Turian et al. (2010)