Amirkabir University of Technology
(Tehran Polytechnic)

Natural Language Processing

Lecture 7: Sparse Word Representation

Amirkabir University of Technology
Dr Momtazi

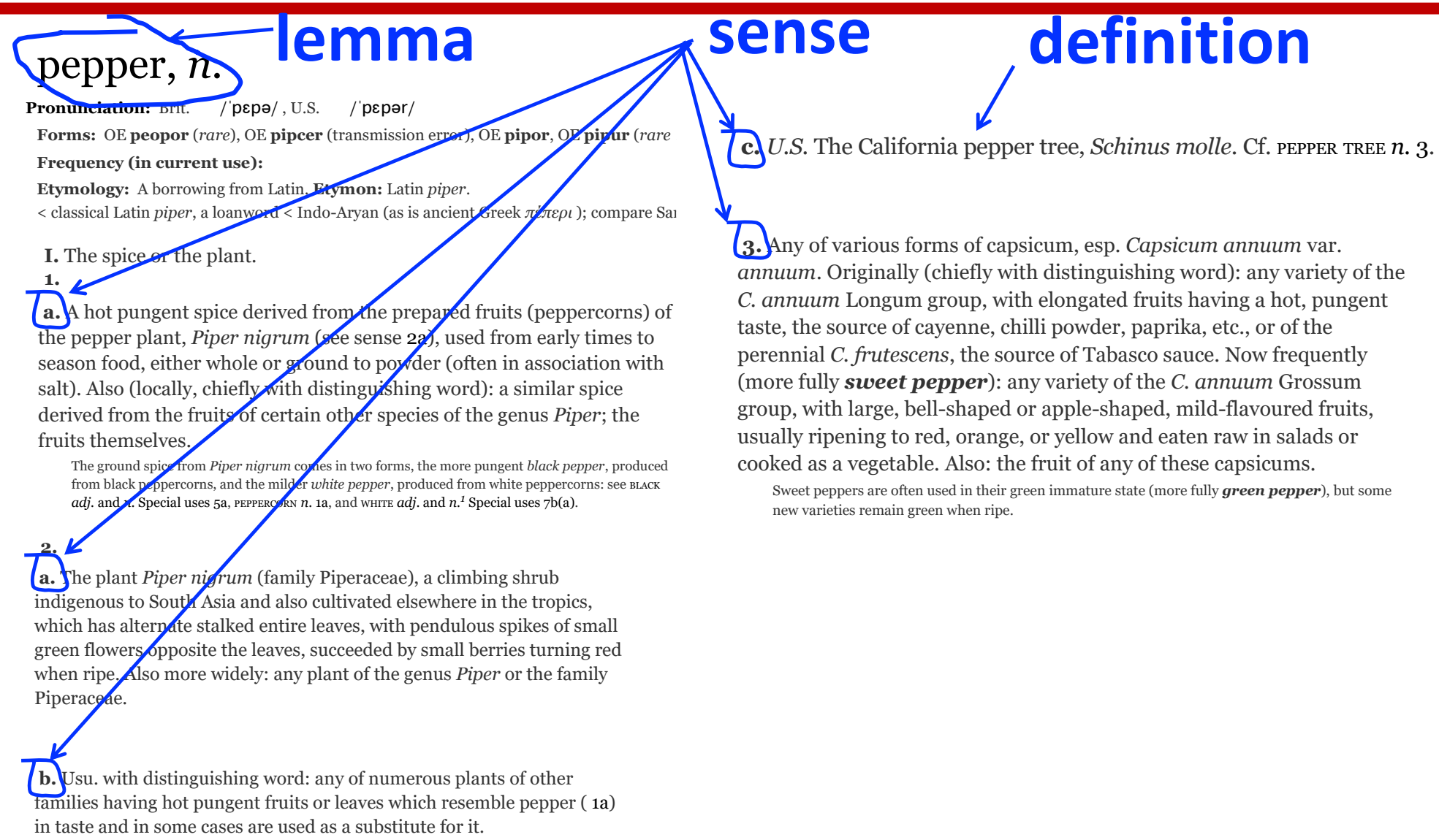
Outline

- **Word Concept**
- Word Vectors

What do words mean?

- First thought: look in a dictionary

Words, Lemmas, Senses, Definitions



Lemma “pepper”

Sense 1: spice from pepper plant

Sense 2: the pepper plant itself

Sense 3: another similar plant (Jamaican pepper)

Sense 4: another plant with peppercorns (California pepper)

Sense 5: *capsicum* (i.e. chili, paprika, bell pepper, etc)

Concept

- A sense or “concept” is the meaning component of a word

- There are relations between senses

Relation: Synonymy

- Synonyms have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - water / H_20

Relation: Synonymy

- Note that there are probably 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.

- The Linguistic Principle of Contrast:
  - Difference in form -> difference in meaning

Relation: Synonymy?

Water/H_20

Big/large

Brave/courageous

Relation: Antonymy

- 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`

Relation: Similarity

- Words with similar meanings.

- Not synonyms, but sharing some element of meaning
  - `car, bicycle`
  - `cow, horse`

Analysis

- Ask humans how similar 2 words are

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)

Relation: Word Relatedness

- Also called "word association"

- Words be related in any way, perhaps via a semantic frame or field
  - `car, bicycle`: **similar**
  - `car, gasoline`: **related**, not similar

Semantic Field

- Words that
  - cover a particular semantic domain
  - bear structured relations with each other.

hospitals

surgeon, scalpel, nurse, anaesthetic, hospital

restaurants

waiter, menu, plate, food, menu, chef

houses

door, roof, kitchen, family, bed

Relation: Superordinate/ Subordinate

- One sense is a **subordinate** of another if the first sense is more specific, denoting a subclass of the other
  - `car` is a subordinate of `vehicle`
  - `mango` is a subordinate of `fruit`

- Conversely **superordinate**
  - `vehicle` is a superordinate of `car`
  - `fruit` is a superordinate of `mango`

Superordinate	vehicle	fruit	furniture
Subordinate	car	mango	chair

Category Levels

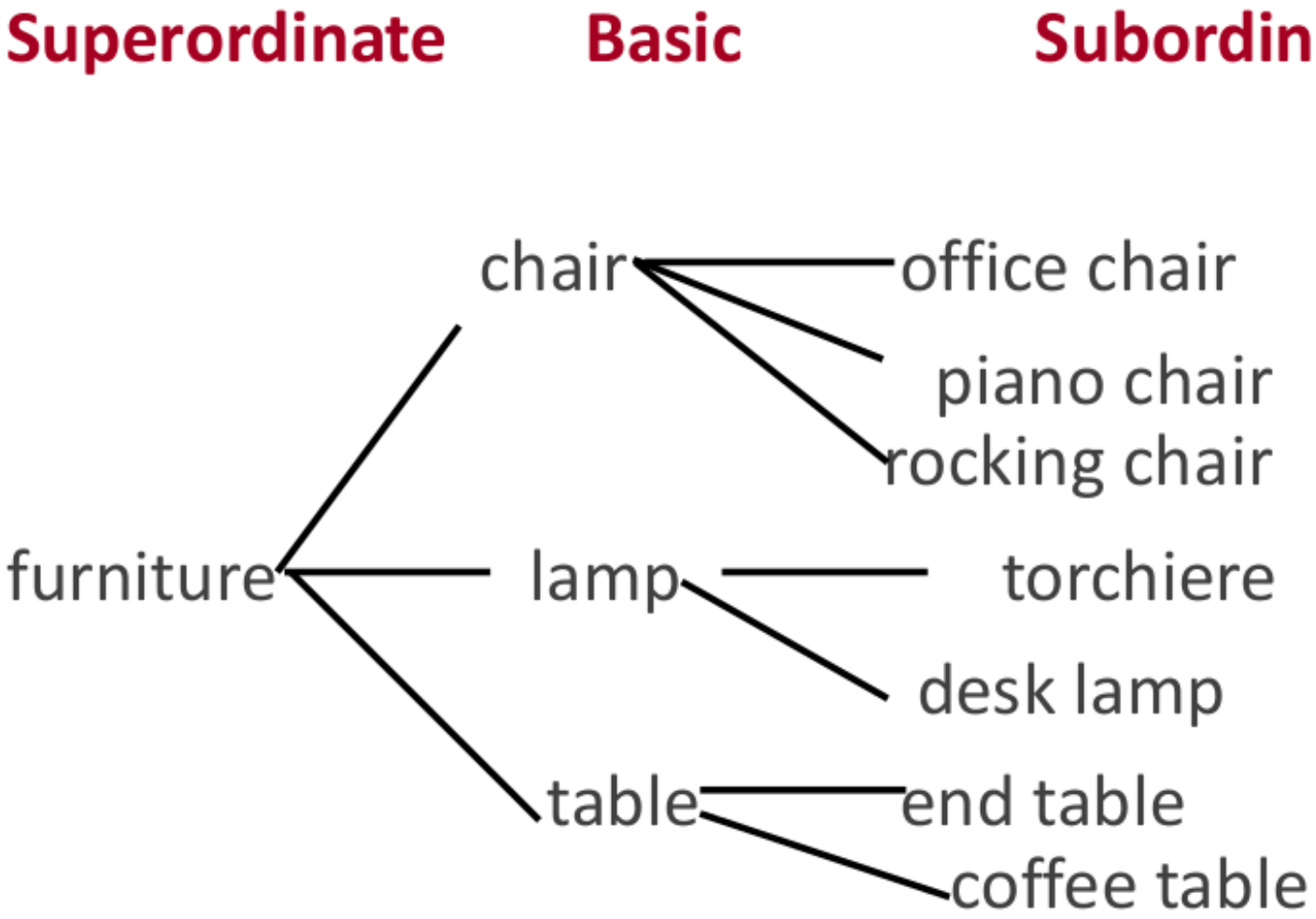
- These levels are not symmetric
- One level of category is distinguished from the others

- Name these items





Category Levels



Cluster of Interactional Properties

- Basic level things are “human-sized”
- Is the level which is learned earliest and at which things are first named
- It is the level at which names are shortest and used most frequently

- Consider chairs
  - We know how to interact with a chair (sitting)
  - Not so clear for superordinate categories like furniture
    - “Imagine a furniture without thinking of a bed/table/chair/specific basic-level category”

So far

- **Concepts** or word senses
  - Have a complex many-to-many association with **words** (homonymy, multiple senses)
- Have relations with each other
  - Synonymy
  - Antonymy
  - Similarity
  - Relatedness
  - Superordinate/subordinate
  - Connotation

- But how to define a concept?

Classical ("Aristotelian") Theory of Concepts

- The meaning of a word:
  - a concept defined by **necessary** and **sufficient** conditions

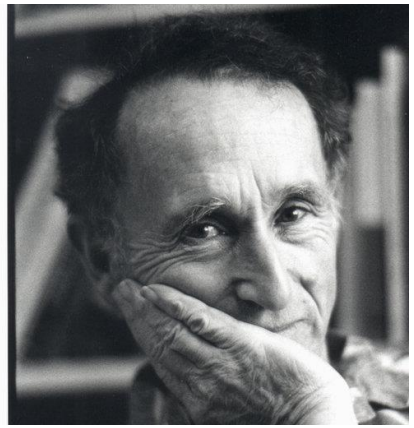
- A **necessary** condition for being an X is a condition C that X must satisfy in order for it to be an X.
  - If not C, then not X
  - "Having four sides" is necessary to be a square.

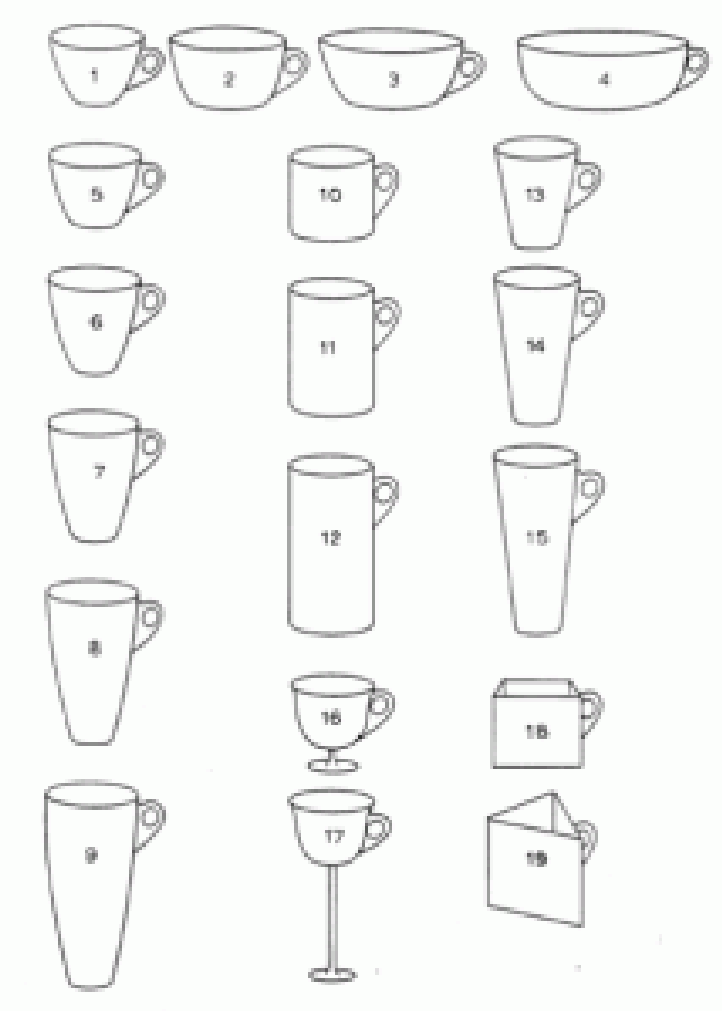
- A **sufficient** condition for being an X is condition such that if something satisfies condition C, then it must be an X.
  - If and only if C, then X

Complex and Context-dependent Features

- Problem 1: The features are complex and may be context-dependent

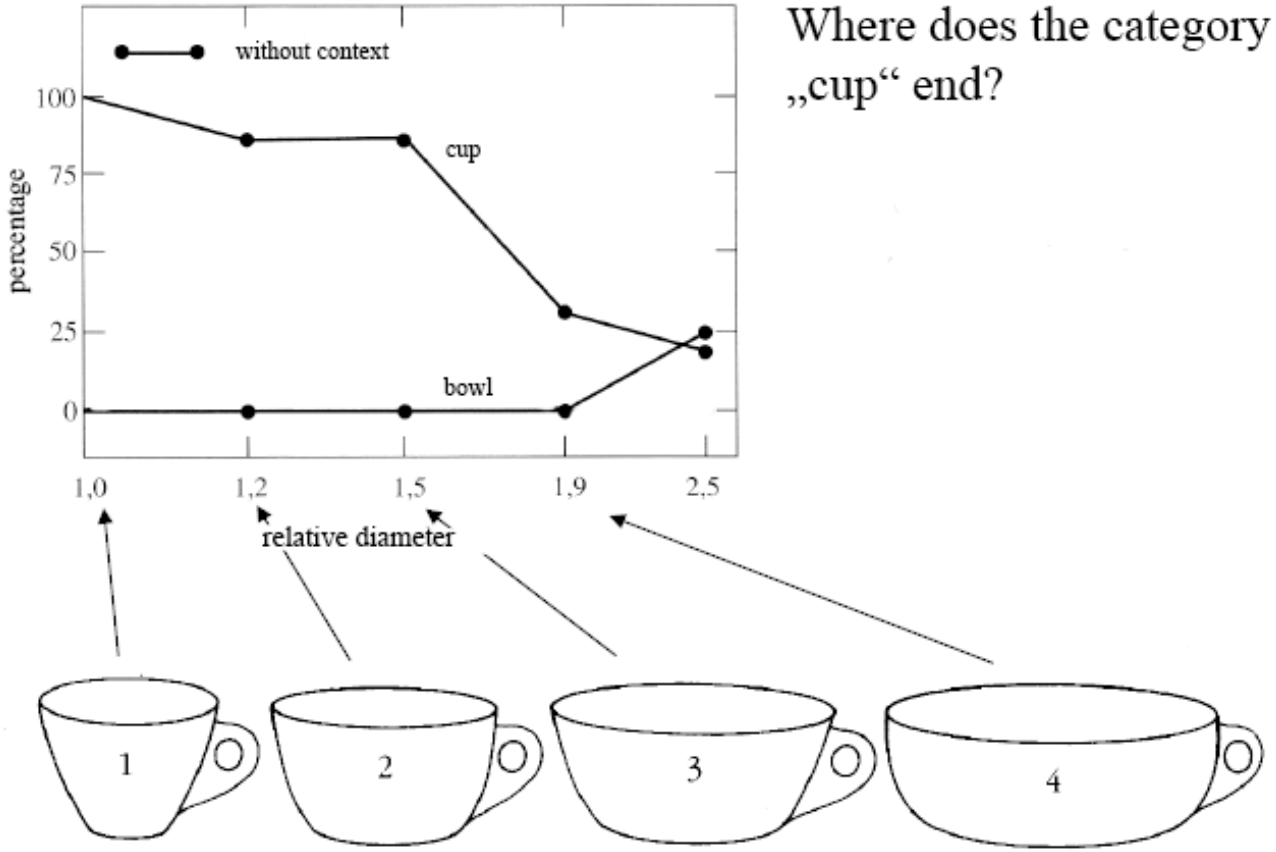
- William Labov. 1975
- What are these?
  - Cup or bowl?





Feature-based Definition

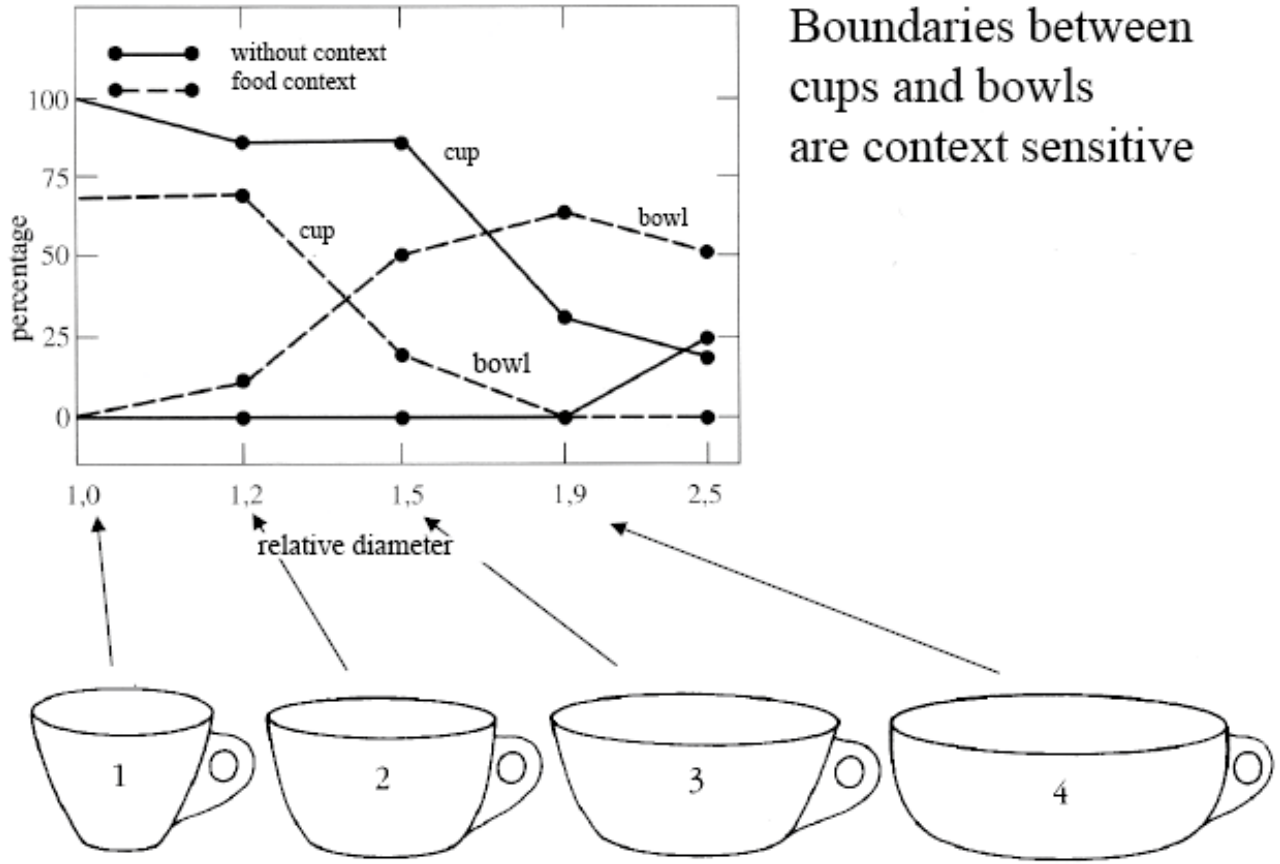
- The category depends on complex features of the object (diameter, etc)



Where does the category „cup“ end?

Context-based Definition

- The category depends on the context! (If there is food in it, it's a bowl)

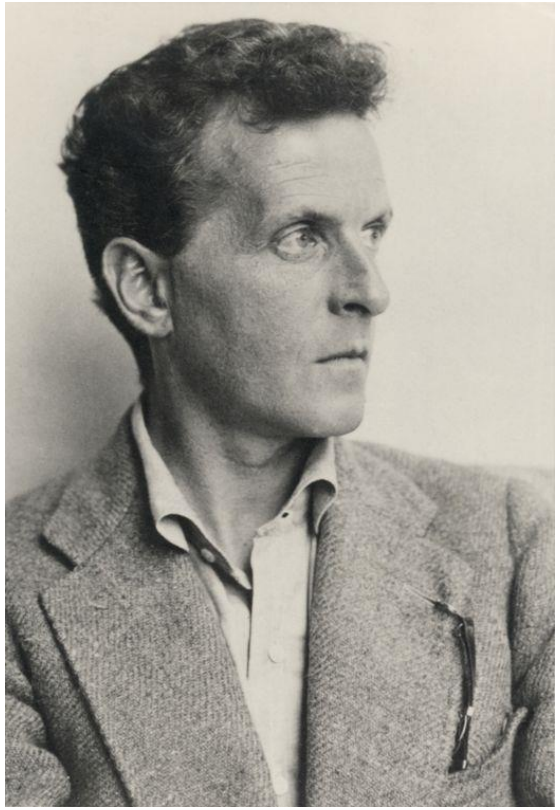


Boundaries between cups and bowls are context sensitive

Ludwig Wittgenstein (1889-1951)

- Philosopher of language
- In his late years, a proponent of studying “ordinary language”

- "The meaning of a word is its use in the language"



Defining Words by their Usages

- In particular, words are defined by their environments (the words around them)

- Zellig Harris (1954): **If A and B have almost identical environments we say that they are synonyms**.

What does “ABC” Means?

- A plate full of *ABC* is on the table
- Everybody likes *ABC*
- *ABC* is cooked within 30 minutes
- We make *ABC* with potato

What does “ABC” Means?

- A plate full of *ABC* is on the table
- Everybody likes *ABC*
- *ABC* is cooked within 30 minutes
- We make *ABC* with potato

⇒ A kind of food

Word Meaning

- We'll build a new model of meaning focusing on similarity
- Each word = a vector
- Similar words are "nearby in space"

Word Vectors

- We define a word as a vector
- Called an "embedding" because it's embedded into a space
- The standard way to represent meaning in NLP
- Fine-grained model of meaning for similarity
  - NLP tasks like sentiment analysis
    - With words, requires **same** word to be in training and test
    - With embeddings: ok if **similar** words occurred!!!
  - Question answering, conversational agents, etc

Outline

- Word Concept
- **Word Vectors**

Word Vectors

- We'll introduce 2 kinds of embeddings
  - Sparse vector representation
    - A common baseline model
    - Words are represented by a simple function of the counts of nearby words
    - Tf-idf
  - Dense vector representation
    - Representation is created by training a model to distinguish nearby and far-away words
    - Word2vec and Glove

Review

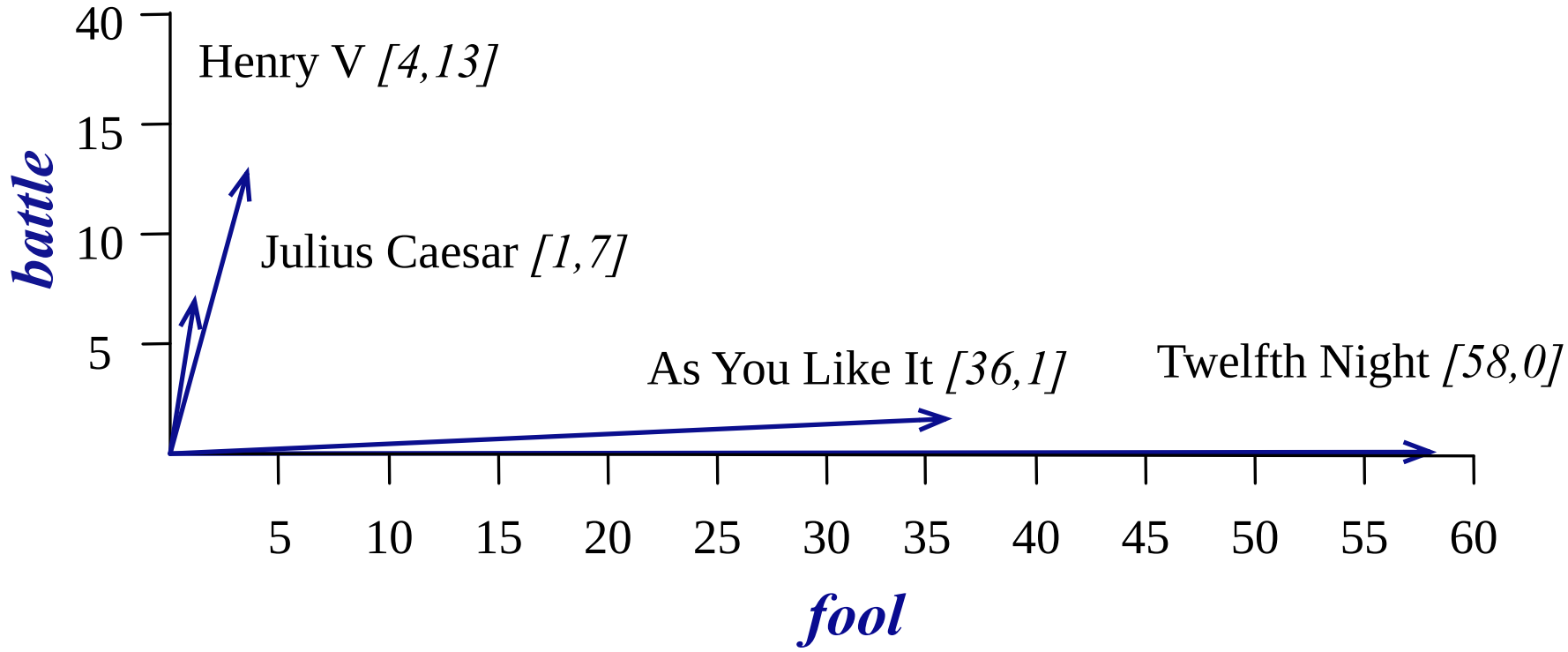
- Words
- Vectors
- Co-occurrence matrices

Term-document Matrix

Each document is represented by a vector of words

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Visualizing Document Vectors



Vectors are the Basis of Information Retrieval

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Vectors are similar for the two comedies
Different than the history

Comedies have more fools and wit and fewer battles.

Words can be Vectors too

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

battle is "the kind of word that occurs in Julius Caesar and Henry V"

fool is "the kind of word that occurs in comedies, especially Twelfth Night"

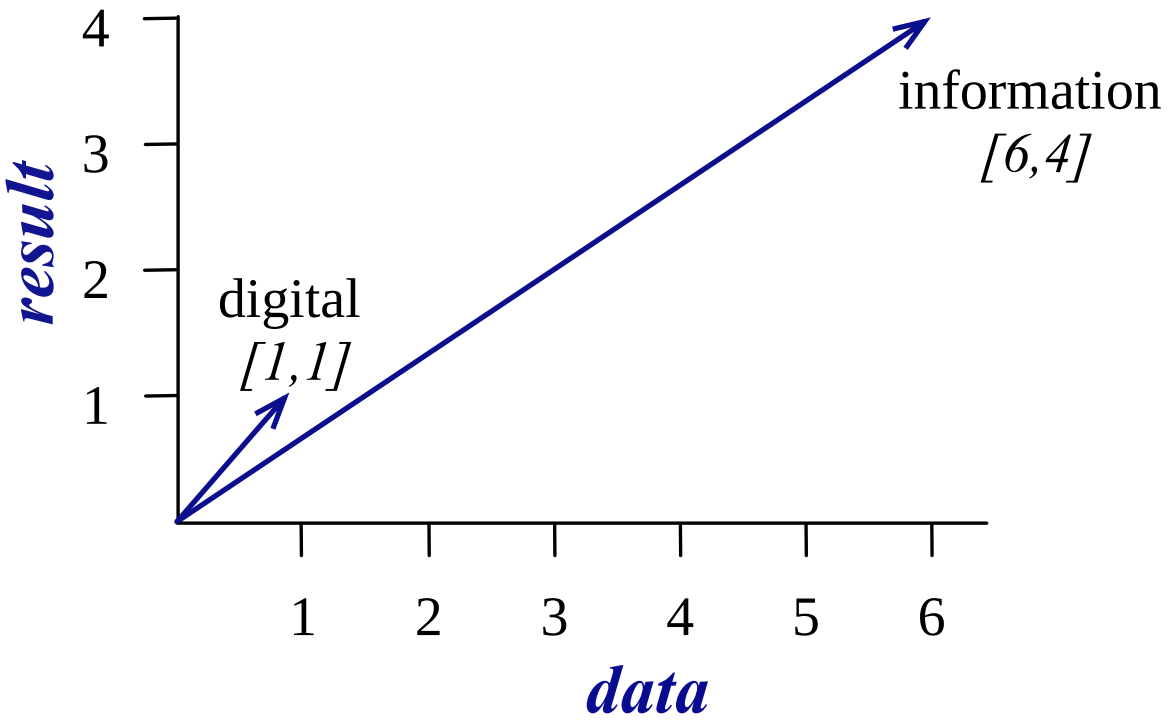
More Common:
Word-Word Matrix (or “Term-Context Matrix")

- Two **words** are similar in meaning if their context vectors are similar

sugar, a sliced lemon, a tablespoonful of	**apricot**	jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

More Common:
Word-Word Matrix (or “Term-Context Matrix")



Cosine for Computing Similarity

- -1: vectors point in opposite directions
- +1: vectors point in same directions
- 0: vectors are orthogonal

- Frequency is non-negative, so cosine range 0-1

$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Cosine for Computing Similarity

- Example

- Which pair of words is more similar?

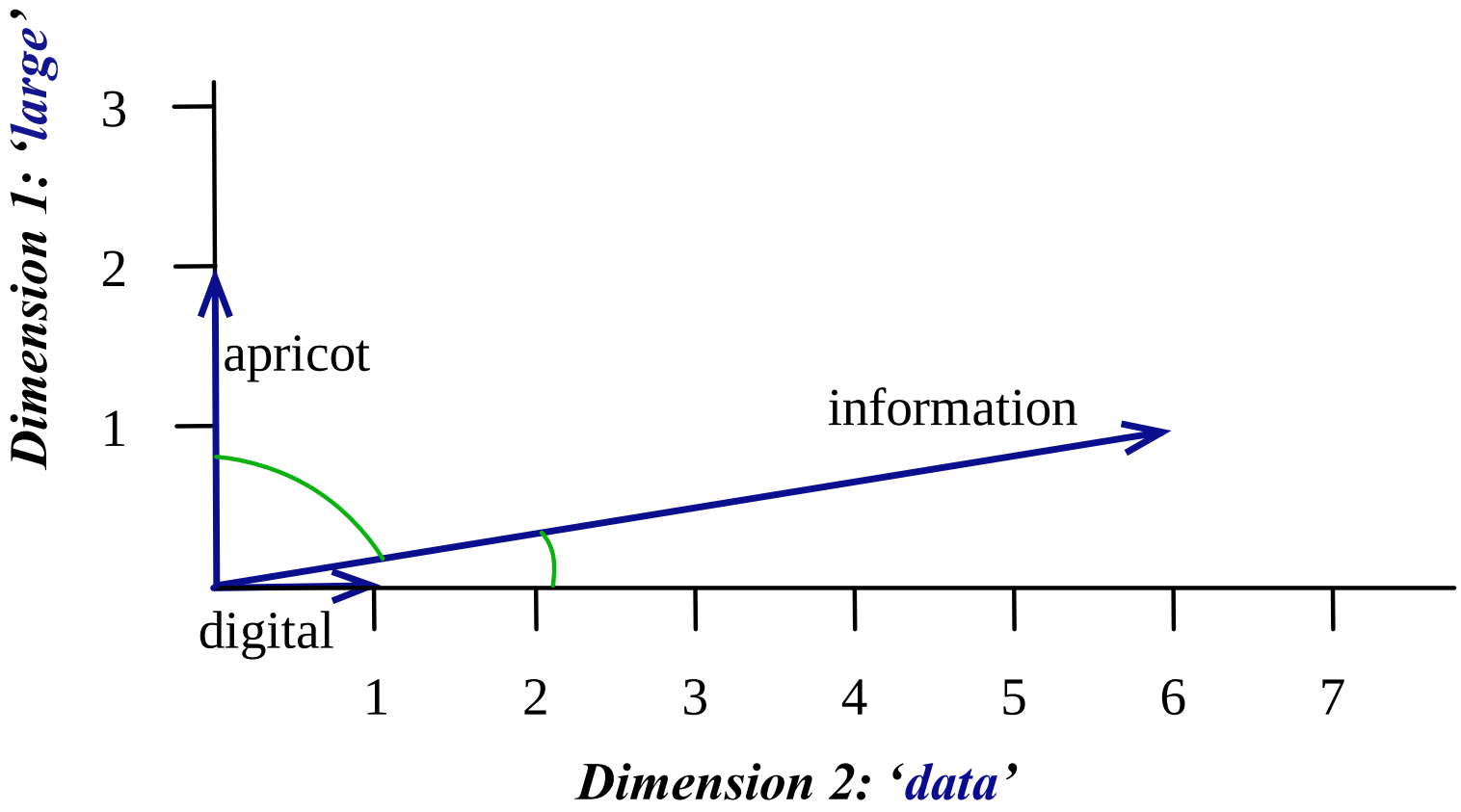
	large	data	computer
apricot	1	0	0
digital	0	1	2
information	1	6	1

cosine(apricot,information) = $\frac{1+0+0}{\sqrt{1+0+0}\sqrt{1+36+1}} = \frac{1}{\sqrt{38}} = .16$

cosine(digital,information) = $\frac{0+6+2}{\sqrt{0+1+4}\sqrt{1+36+1}} = \frac{8}{\sqrt{38}\sqrt{5}} = .58$

cosine(apricot,digital) = $\frac{0+0+0}{\sqrt{1+0+0}\sqrt{0+1+4}} = 0$

Visualizing Cosines



Raw Frequency

- But raw frequency is a bad representation
- Frequency is clearly useful; if *sugar* appears a lot near *apricot*, that's useful information.
- But overly frequent words like *the*, *it,* or *they* are not very informative about the context
- Need a function that resolves this frequency paradox!

tf-idf: Combine Two Factors

- tf: term frequency. frequency count (usually log-transformed):

$$\mathrm{tf}_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{count}(t,d) & \text{if } \mathrm{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Idf: inverse document frequency:

$$\mathrm{idf}_i = \log\left(\frac{N}{\mathrm{df}_i}\right)$$

Total # of docs in collection

# of docs that have word i

- tf-idf value for word t in document d:

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

Summary: tf-idf

- Compare two words using tf-idf cosine to see if they are similar
- Compare two documents
  - Take the centroid of vectors of all the words in the document
  - Centroid document vector is:

$$d = \frac{w_1 + w_2 + \ldots + w_k}{k}$$

Positive Pointwise Mutual Information (PPMI)

- An alternative to tf-idf
- Ask whether a context word is **particularly informative** about the target word.

Pointwise Mutual Information

- Pointwise mutual information:
  - Do events x and y co-occur more than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- PMI between two words:
  - Do words x and y co-occur more than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

Positive Pointwise Mutual Information

- PMI ranges from $-\infty$ to $+\infty$
- But the negative values are problematic
  - Things are co-occurring **less than** we expect by chance
  - Unreliable without enormous corpora
    - Imagine w1 and w2 whose probability is each 10^{-6}
    - Hard to be sure p(w1,w2) is significantly different than 10^{-12}
  - Plus it's not clear people are good at "unrelatedness"

- So we just replace negative PMI values by 0
- Positive PMI (PPMI) between word1 and word2:

$$\text{PPMI}(word_1, word_2) = \max\left(\log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}, 0\right)$$

Comparison

Count(w,context)

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

p(w,context)

	computer	data	pinch	result	sugar	**p(w)**
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

PPMI(w,context)

	computer	data	pinch	result	sugar
apricot	-	-	2.25	-	2.25
pineapple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-

Weighting PMI

- PMI is biased toward infrequent events
  - Very rare words have very high PMI values

- Two solutions:
  - Give rare words slightly higher probabilities
  - Use add-one smoothing (which has a similar effect)

Summary

- Survey of Lexical Semantics
- Idea of embeddings: Represent a word as a function of its distribution with other words
  - Tf-idf
  - PPMI

- Next lecture: sparse vs dense embeddings, word2vec

Further Reading

- Speech and Language Processing (3rd ed. draft)
  - Chapter 6