CSCI 5832
Natural Language Processing

Jim Martin

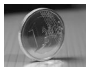
Lecture 6

Today 1/31

- Probability
  - Basic probability
  - Conditional probability
  - Bayes Rule
- Language Modeling (N-grams)
  - N-gram Intro
  - The Chain Rule
  - Smoothing: Add-1

Probability Basics

- Experiment (trial)
  - Repeatable procedure with well-defined possible outcomes
- Sample Space (S)
  - the set of all possible outcomes
  - *finite or infinite*
  - Example
    - coin toss experiment
    - possible outcomes: S = {heads, tails}
  - Example
    - die toss experiment
    - possible outcomes: S = {1,2,3,4,5,6}



Slides from Sandiway Fong

Probability Basics

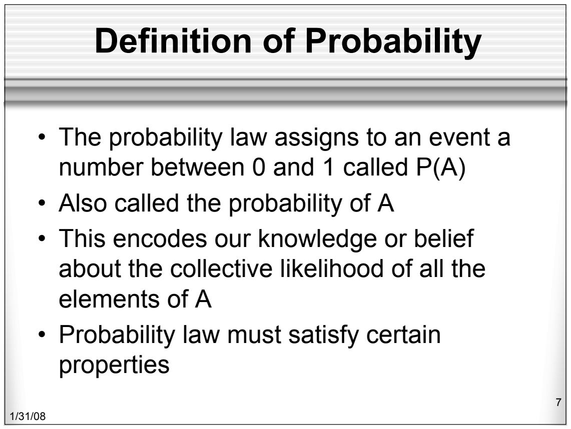
- Definition of sample space depends on what we are asking
  - Sample Space (S): the set of all possible outcomes
  - Example
    - die toss experiment for whether the number is even or odd
    - possible outcomes: {even,odd}
    - *not* {1,2,3,4,5,6}

More Definitions

- Events
  - an ***event*** is any subset of outcomes from the ***sample space***
- Example
  - Die toss experiment
    - Let A represent the event such that the outcome of the die toss experiment is divisible by 3
    - A = {3,6}
    - A is a subset of the sample space S= {1,2,3,4,5,6}
- Example
  - Draw a card from a deck
    - suppose sample space S = {heart,spade,club,diamond} (*four suits*)
    - let A represent the event of drawing a heart
    - let B represent the event of drawing a red card
    - A = {heart}
    - B = {heart,diamond}

Probability Basics

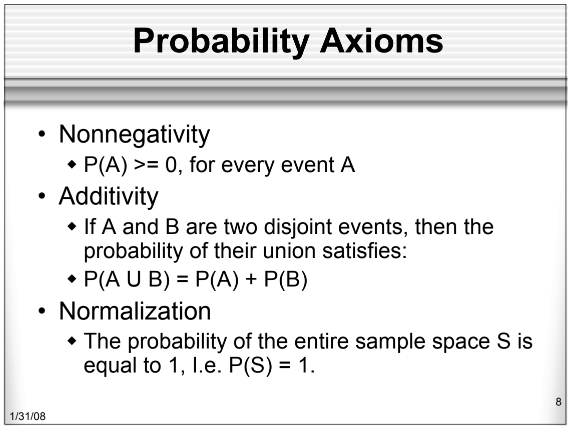
- Some definitions
  - Counting
    - suppose operation o_i can be performed in n_i ways, then
    - a sequence of k operations $o_1 o_2 \ldots o_k$
    - can be performed in $n_1 \times n_2 \times \ldots \times n_k$ ways
  - Example
    - die toss experiment, 6 possible outcomes
    - two dice are thrown at the same time
    - number of sample points in sample space = 6 × 6 = 36



Definition of Probability

- The probability law assigns to an event a number between 0 and 1 called P(A)
- Also called the probability of A
- This encodes our knowledge or belief about the collective likelihood of all the elements of A
- Probability law must satisfy certain properties

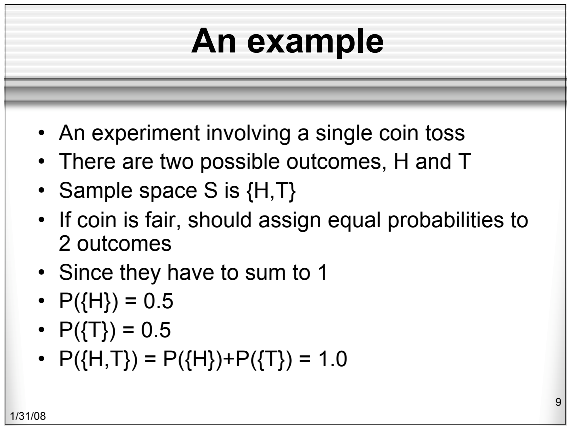
1/31/08 7



Probability Axioms

- Nonnegativity
  - P(A) >= 0, for every event A
- Additivity
  - If A and B are two disjoint events, then the probability of their union satisfies:
  - P(A U B) = P(A) + P(B)
- Normalization
  - The probability of the entire sample space S is equal to 1, I.e. P(S) = 1.

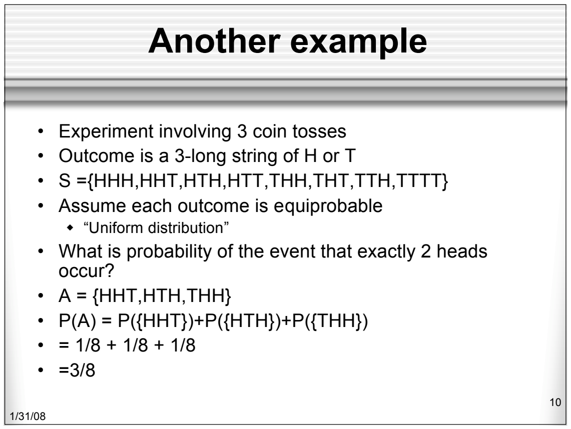
1/31/08 8



An example

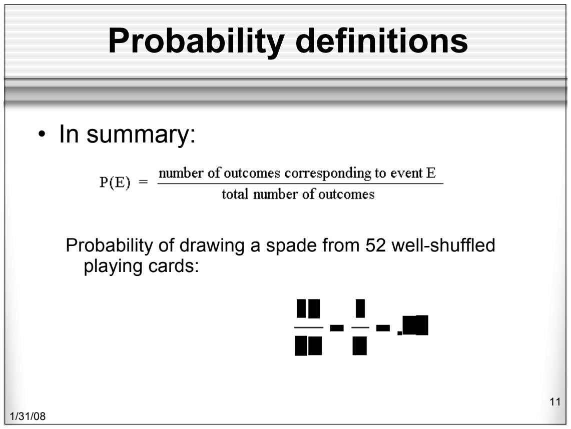
- An experiment involving a single coin toss
- There are two possible outcomes, H and T
- Sample space S is {H,T}
- If coin is fair, should assign equal probabilities to 2 outcomes
- Since they have to sum to 1
- P({H}) = 0.5
- P({T}) = 0.5
- P({H,T}) = P({H})+P({T}) = 1.0

1/31/08 9



Another example

- Experiment involving 3 coin tosses
- Outcome is a 3-long string of H or T
- S ={HHH,HHT,HTH,HTT,THH,THT,TTH,TTTT}
- Assume each outcome is equiprobable
  - “Uniform distribution”
- What is probability of the event that exactly 2 heads occur?
- A = {HHT,HTH,THH}
- P(A) = P({HHT})+P({HTH})+P({THH})
- = 1/8 + 1/8 + 1/8
- =3/8



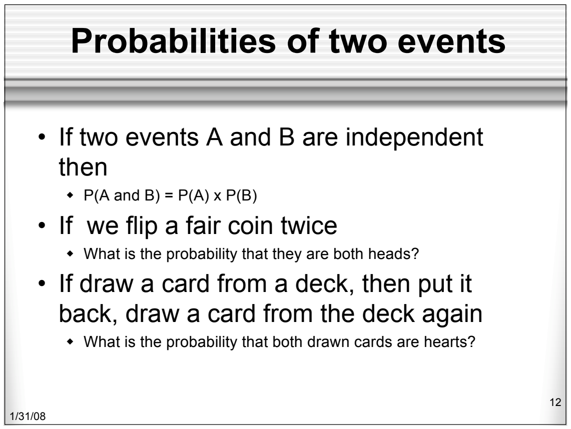
Probability definitions

- In summary:

$$P(E) = \frac{\text{number of outcomes corresponding to event E}}{\text{total number of outcomes}}$$

Probability of drawing a spade from 52 well-shuffled playing cards:

[illegible]



Probabilities of two events

- If two events A and B are independent then
  - P(A and B) = P(A) x P(B)
- If we flip a fair coin twice
  - What is the probability that they are both heads?
- If draw a card from a deck, then put it back, draw a card from the deck again
  - What is the probability that both drawn cards are hearts?

How about non-uniform probabilities?

- A biased coin,
  - twice as likely to come up tails as heads,
  - is tossed twice
- What is the probability that at least one head occurs?
- Sample space = {hh, ht, th, tt}
- Sample points/probability for the event:
  - ht 1/3 x 2/3 = **2/9** hh 1/3 x 1/3= **1/9**
  - th 2/3 x 1/3 = **2/9** tt 2/3 x 2/3 = 4/9
- Answer: 5/9 = ≈0.56 (*sum of weights in **red***)

Moving toward language

- What's the probability of drawing a 2 from a deck of 52 cards with four 2s?

[illegible]

- What's the probability of a random word (from a random dictionary page) being a verb?

[illegible]

Probability and part of speech tags

- What's the probability of a random word (from a random dictionary page) being a verb?

[illegible]

- How to compute each of these
- All words = just count all the words in the dictionary
- # of ways to get a verb: number of words which are verbs!
- If a dictionary has 50,000 entries, and 10,000 are verbs…. P(V) is 10000/50000 = 1/5 = .20

Conditional Probability

- A way to reason about the outcome of an experiment based on partial information
  - In a word guessing game the first letter for the word is a “t”. What is the likelihood that the second letter is an “h”?
  - How likely is it that a person has a disease given that a medical test was negative?
  - A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?

More precisely

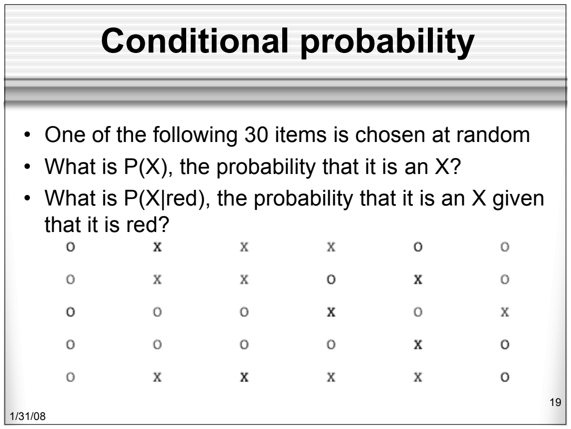
- Given an experiment, a corresponding sample space S, and a probability law
- Suppose we know that the outcome is within some given event B
- We want to quantify the likelihood that the outcome also belongs to some other given event A.
- We need a new probability law that gives us the conditional probability of A given B
- P(A|B)

An intuition

- A is “it’s snowing now”.
- P(A) in normally arid Colorado is .01
- B is “it was snowing ten minutes ago”

- P(A|B) means “what is the probability of it snowing now if it was snowing 10 minutes ago”
- P(A|B) is probably way higher than P(A)
- Perhaps P(A|B) is .10

- Intuition: The knowledge about B should change (update) our estimate of the probability of A.



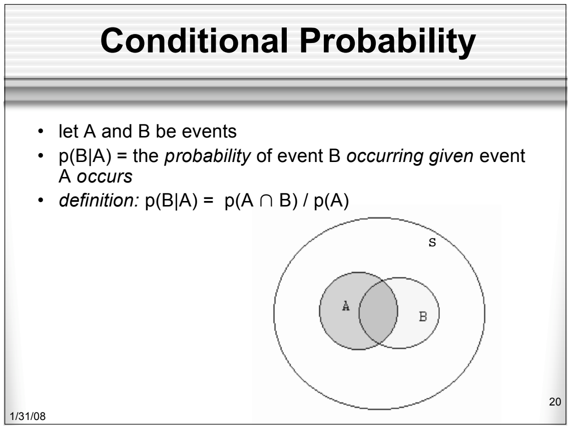
Conditional probability

- One of the following 30 items is chosen at random
- What is P(X), the probability that it is an X?
- What is P(X|red), the probability that it is an X given that it is red?

o	x	x	x	o	o
o	x	x	o	x	o
o	o	o	x	o	x
o	o	o	o	x	o
o	x	x	x	x	o

1/31/08

19

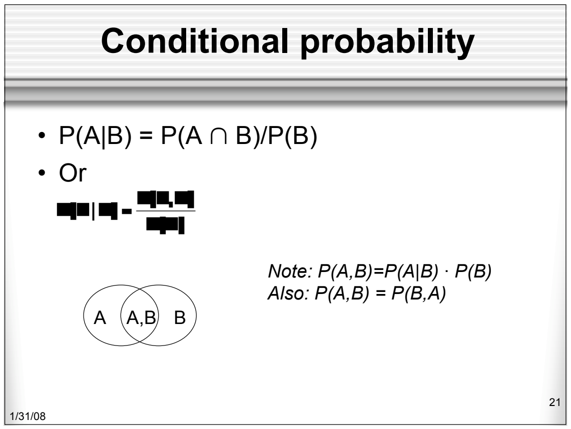


Conditional Probability

- let A and B be events
- p(B|A) = the *probability* of event B *occurring given* event A *occurs*
- *definition:* p(B|A) = p(A ∩ B) / p(A)

1/31/08

20



Conditional probability

- P(A|B) = P(A ∩ B)/P(B)
- Or

[illegible]

Note: P(A,B)=P(A|B) · P(B)
Also: P(A,B) = P(B,A)

1/31/08

21

Independence

- What is P(A,B) if A and B are independent?
- P(A,B)=P(A) · P(B) iff A,B independent.

P(heads,tails) = P(heads) · P(tails) = .5 · .5 = .25

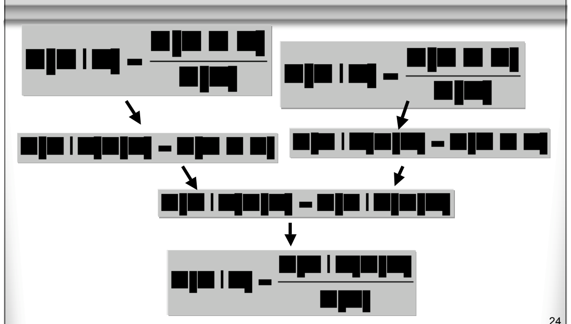
Note: P(A|B)=P(A) iff A,B independent
Also: P(B|A)=P(B) iff A,B independent

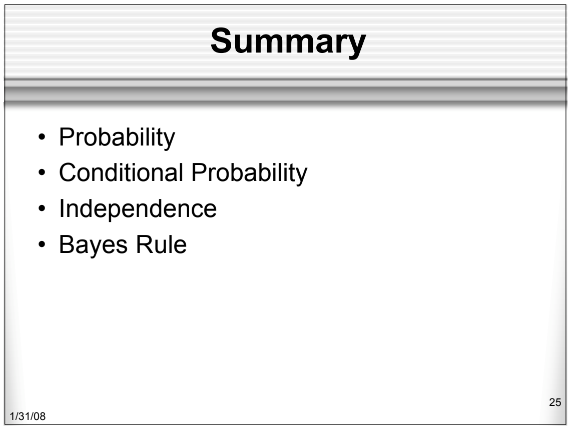
Bayes Theorem

[illegible]

- Swap the conditioning
- Sometimes easier to estimate one kind of dependence than the other

Deriving Bayes Rule

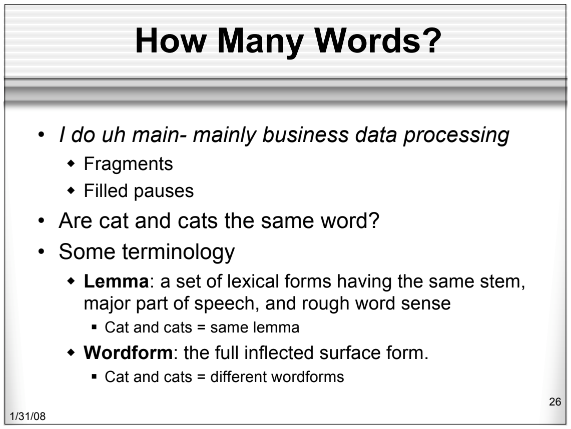




Summary

- Probability
- Conditional Probability
- Independence
- Bayes Rule

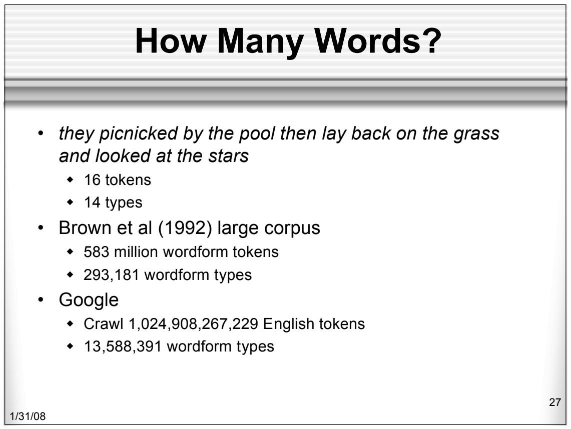
1/31/08 25



How Many Words?

- *I do uh main- mainly business data processing*
  - Fragments
  - Filled pauses
- Are cat and cats the same word?
- Some terminology
  - **Lemma**: a set of lexical forms having the same stem, major part of speech, and rough word sense
    - Cat and cats = same lemma
  - **Wordform**: the full inflected surface form.
    - Cat and cats = different wordforms

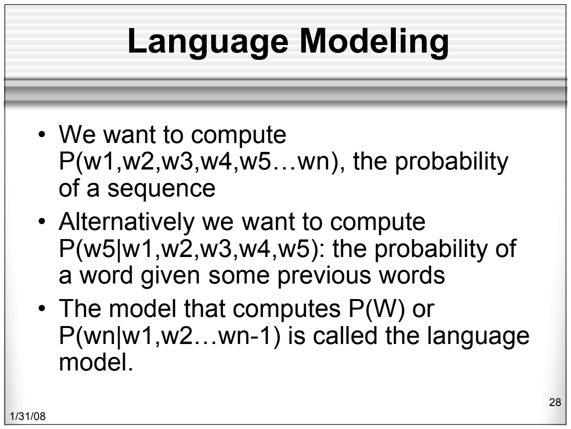
1/31/08 26



How Many Words?

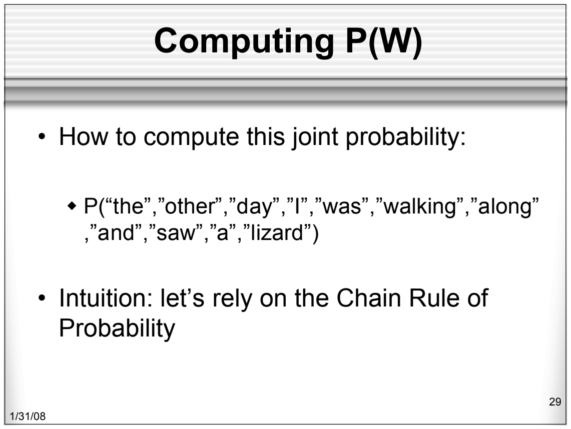
- *they picnicked by the pool then lay back on the grass and looked at the stars*
  - 16 tokens
  - 14 types
- Brown et al (1992) large corpus
  - 583 million wordform tokens
  - 293,181 wordform types
- Google
  - Crawl 1,024,908,267,229 English tokens
  - 13,588,391 wordform types

1/31/08 27



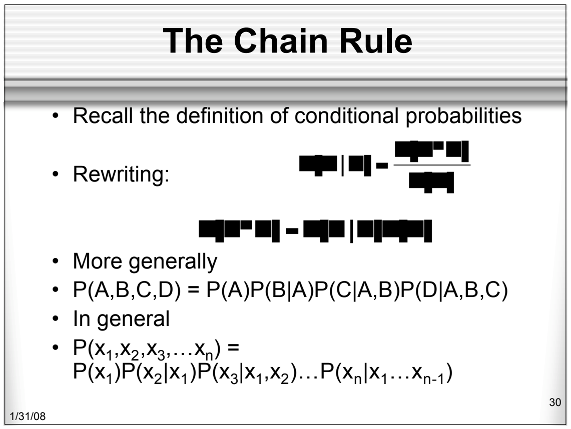
Language Modeling

- We want to compute P(w1,w2,w3,w4,w5…wn), the probability of a sequence
- Alternatively we want to compute P(w5|w1,w2,w3,w4,w5): the probability of a word given some previous words
- The model that computes P(W) or P(wn|w1,w2…wn-1) is called the language model.



Computing P(W)

- How to compute this joint probability:
  - P("the","other","day","I","was","walking","along" ,"and","saw","a","lizard")
- Intuition: let's rely on the Chain Rule of Probability

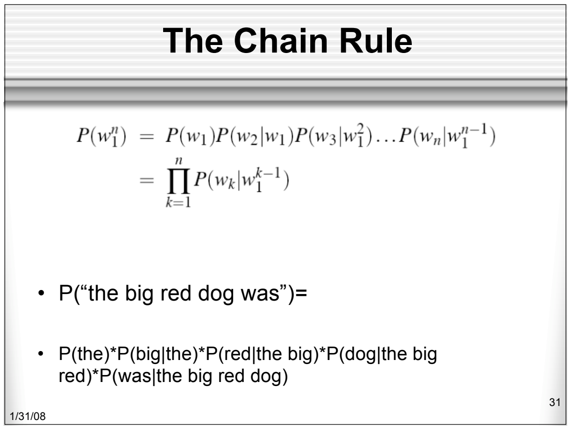


The Chain Rule

- Recall the definition of conditional probabilities
- Rewriting: [illegible]

[illegible]

- More generally
- P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)
- In general
- $P(x_1,x_2,x_3,…x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)…P(x_n|x_1…x_{n-1})$

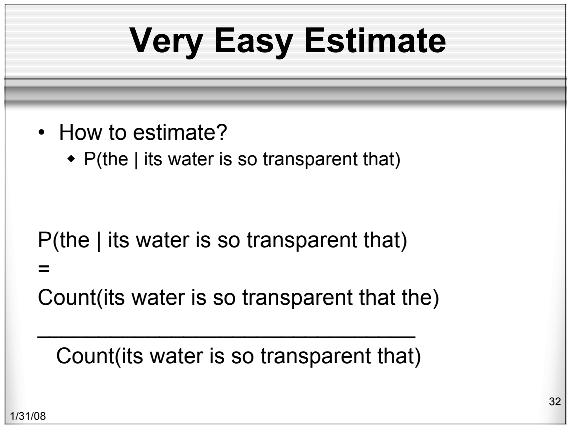


The Chain Rule

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned}$$

- P(“the big red dog was”)=

- P(the)*P(big|the)*P(red|the big)*P(dog|the big red)*P(was|the big red dog)

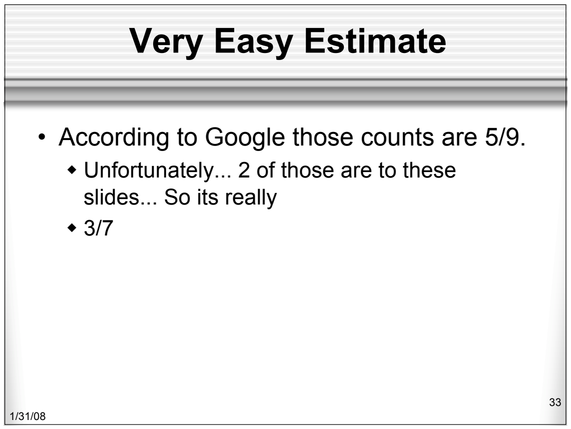


Very Easy Estimate

- How to estimate?
  - P(the | its water is so transparent that)

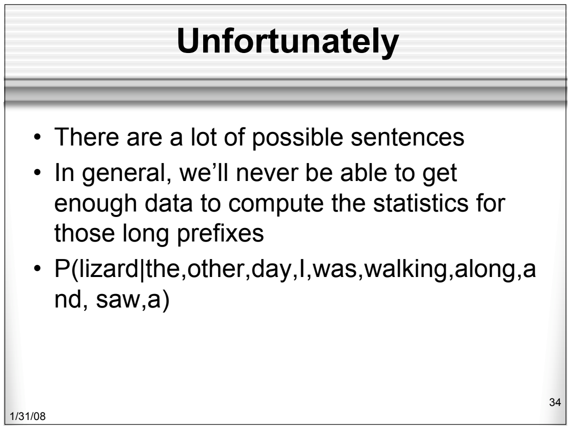
P(the | its water is so transparent that)
=

$$\frac{\text{Count(its water is so transparent that the)}}{\text{Count(its water is so transparent that)}}$$



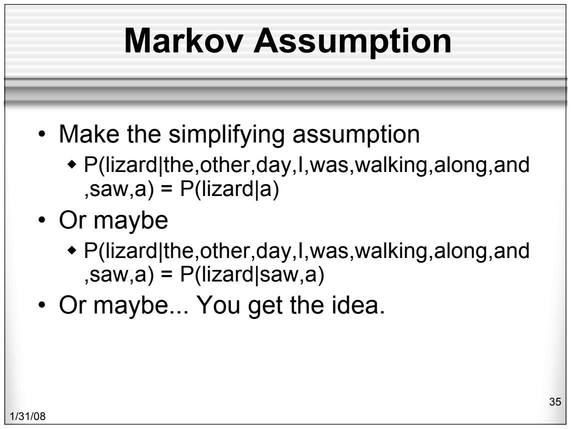
Very Easy Estimate

- According to Google those counts are 5/9.
  - Unfortunately... 2 of those are to these slides... So its really
  - 3/7



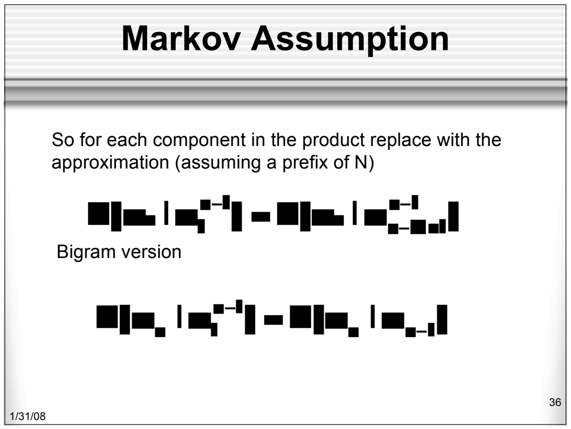
Unfortunately

- There are a lot of possible sentences
- In general, we'll never be able to get enough data to compute the statistics for those long prefixes
- P(lizard|the,other,day,I,was,walking,along,a nd, saw,a)



Markov Assumption

- Make the simplifying assumption
  - P(lizard|the,other,day,I,was,walking,along,and ,saw,a) = P(lizard|a)
- Or maybe
  - P(lizard|the,other,day,I,was,walking,along,and ,saw,a) = P(lizard|saw,a)
- Or maybe... You get the idea.



Markov Assumption

So for each component in the product replace with the approximation (assuming a prefix of N)

[illegible]

Bigram version

[illegible]

Estimating bigram probabilities

- The Maximum Likelihood Estimate

[illegible]

[illegible]

An example

- <s> I am Sam </s>
- <s> Sam I am </s>
- <s> I do not like green eggs and ham </s>

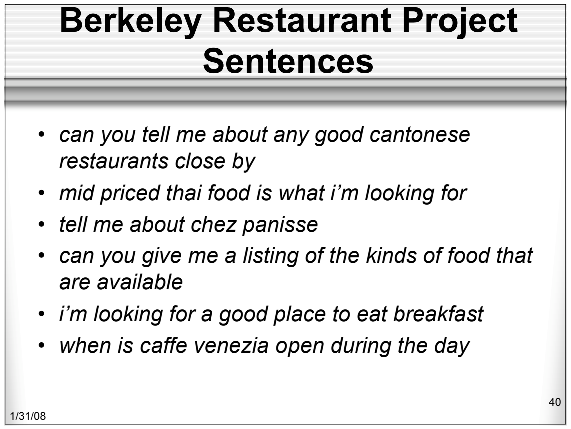
$P(\texttt{I}\,|\,\texttt{<s>}) = \frac{2}{3} = .67 \qquad P(\texttt{Sam}\,|\,\texttt{<s>}) = \frac{1}{3} = .33 \qquad P(\texttt{am}\,|\,\texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>}\,|\,\texttt{Sam}) = \frac{1}{2} = 0.5 \qquad P(\texttt{Sam}\,|\,\texttt{am}) = \frac{1}{2} = .5 \qquad P(\texttt{do}\,|\,\texttt{I}) = \frac{1}{3} = .33$

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

Maximum Likelihood Estimates

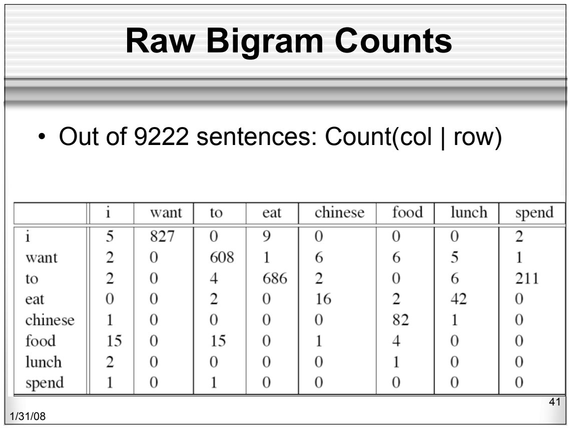
- The maximum likelihood estimate of some parameter of a model M from a training set T
  - Is the estimate that maximizes the likelihood of the training set T given the model M
- Suppose the word Chinese occurs 400 times in a corpus of a million words (Brown corpus)
- What is the probability that a random word from some other text from the same distribution will be "Chinese"
- MLE estimate is 400/1000000 = .004
  - This may be a bad estimate for some other corpus
- But it is the **estimate** that makes it **most likely** that "Chinese" will occur 400 times in a million word corpus.



Berkeley Restaurant Project Sentences

- *can you tell me about any good cantonese restaurants close by*
- *mid priced thai food is what i’m looking for*
- *tell me about chez panisse*
- *can you give me a listing of the kinds of food that are available*
- *i’m looking for a good place to eat breakfast*
- *when is caffe venezia open during the day*

1/31/08 40

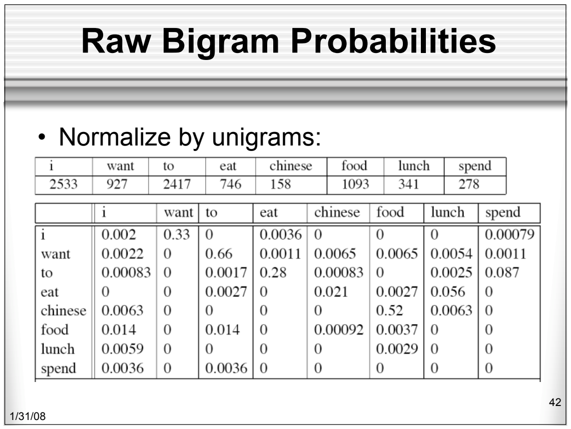


Raw Bigram Counts

- Out of 9222 sentences: Count(col | row)

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

1/31/08 41



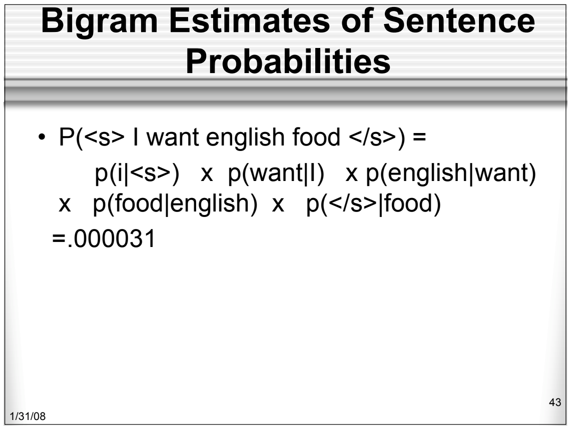
Raw Bigram Probabilities

- Normalize by unigrams:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

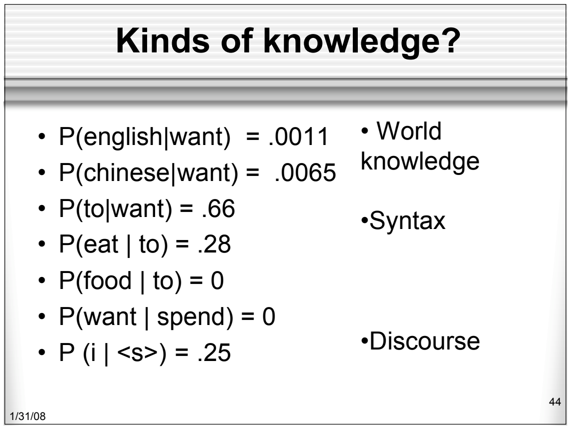
	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

1/31/08 42



Bigram Estimates of Sentence Probabilities

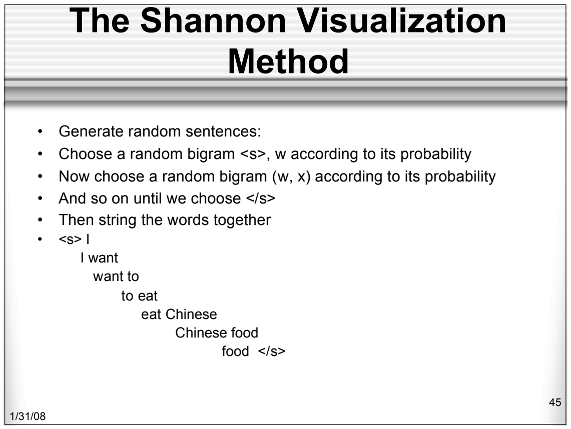
- P(<s> I want english food </s>) =
  p(i|<s>) x p(want|I) x p(english|want)
  x p(food|english) x p(</s>|food)
  =.000031



Kinds of knowledge?

- P(english|want) = .0011
- P(chinese|want) = .0065
- P(to|want) = .66
- P(eat | to) = .28
- P(food | to) = 0
- P(want | spend) = 0
- P (i | <s>) = .25

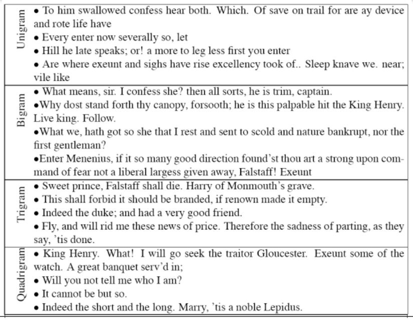
- World knowledge
- Syntax
- Discourse



The Shannon Visualization Method

- Generate random sentences:
- Choose a random bigram <s>, w according to its probability
- Now choose a random bigram (w, x) according to its probability
- And so on until we choose </s>
- Then string the words together
- <s> I
  I want
  want to
  to eat
  eat Chinese
  Chinese food
  food </s>

Shakespeare

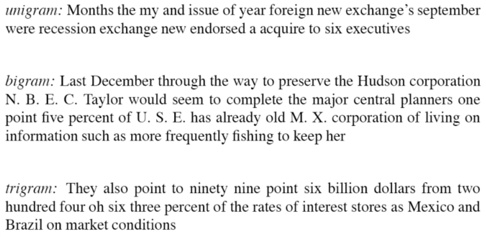


Unigram	• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have • Every enter now severally so, let • Hill he late speaks; or! a more to leg less first you enter • Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like
Bigram	• What means, sir. I confess she? then all sorts, he is trim, captain. •Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. •What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman? •Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
Trigram	• Sweet prince, Falstaff shall die. Harry of Monmouth's grave. • This shall forbid it should be branded, if renown made it empty. • Indeed the duke; and had a very good friend. • Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
Quadrigram	• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; • Will you not tell me who I am? • It cannot be but so. • Indeed the short and the long. Marry, 'tis a noble Lepidus.

Shakespeare as corpus

- N=884,647 tokens, V=29,066
- Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams: so, 99.96% of the possible bigrams were never seen (have zero entries in the table)
- Quadrigrams worse: What's coming out looks like Shakespeare because it ***is*** Shakespeare

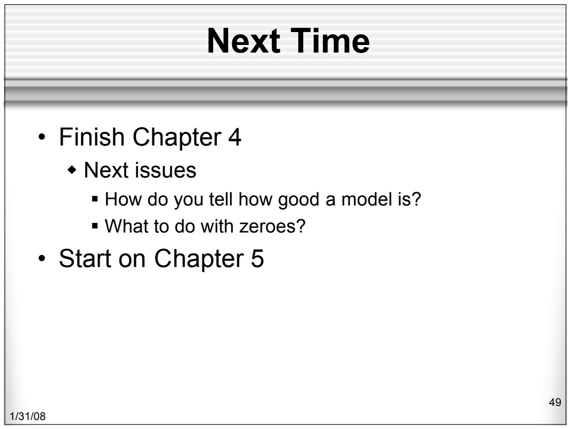
The Wall Street Journal is Not Shakespeare



unigram: Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

bigram: Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

trigram: They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions



Next Time

- Finish Chapter 4
  - Next issues
    - How do you tell how good a model is?
    - What to do with zeroes?
- Start on Chapter 5

1/31/08

49