Linguistics 696f
Hammond
Spring ’03

# N-gram models

## A. Overview

(1) a. Unigrams
b. Why do this?
c. Bigrams
d. N-gram approximation
e. Problems
f. Software

## B. Simple unigrams

(2) Peter Piper picked a peck of pickled pepper.
Where’s the pickled pepper that Peter Piper picked?

(3) $\text{P(Peter)} = \frac{2}{16} = .125$

## C. Why do this?

(4) Speech recognition:
Given some ambiguous signal, is it *Peter* or *peck*?

(5) If P(Peter) = .125 in one text, but P(Peter) = .625 in another, are the texts:
a. by the same author?
b. on the same topic?
c. in the same language?

## D. Two problems

(6) Peter Peter Peter. . .

(7) Compare the probabilities of these two texts:
a. Peter Piper picked
b. picked Piper Peter

(8) $\text{P(Peter)} \times \text{P(Piper)} \times \text{P(picked)} = \text{P(picked)} \times \text{P(Piper)} \times \text{P(Peter)}$

## E. Bigrams

(9) $\text{P}(w_1 w_2 \ldots w_i) = \text{P}(w_1) \times \text{P}(w_2|w_1) \times \ldots \times \text{P}(w_i|w_{i-1})$

(10) $\text{P}(w_i|w_{i-1}) = \frac{|w_{i-1}w_i|}{|w_{i-1}|}$

(11) $\text{P(Peter)} \times \text{P(Piper|Peter)} \times \text{P(picked|Piper)} = ?$
$\text{P(picked)} \times \text{P(Piper|picked)} \times \text{P(Peter|Piper)} = ?$

## F. Two more problems

(12) Calculating bigrams like this does *not* result in a probability distribution. Why?

(13) Calculating the probability of a text in terms of bigrams is *not* an instance of the Chain Rule. Why?

## G. N-gram approximation

(14) "White Fang": Jack London

(15) a. so her they dog no but there with in so
b. as not him they so he a that away then
c. be when dogs then up there he fang by a
d. on dogs out his and out he the away out
e. they then that on his into upon been their she
f. fang him this up dogs were he dogs no
g. by fang to into when him their when upon
h. up them at the was a been with there down
i. then down be him and on time one as into
j. as them be to for were that his at when

(16) a. half feet on everywhere upon itself as strongest dog
b. far outweighed a hostile movement beside scott you know
c. judge unknown was because it toward personal life
d. everybody gave himself to cheapen himself off with
e. it bristled fiercely belligerent and save once and death
f. because they spoke matt should be used his tail
g. turn 'm time i counted the horse up live
h. beast that cautiously it discovered an act of plenty
i. fatty's gone before had thought in matt argued stubbornly
j. what evil that night was flying brands from weakness

## H. Software

(17) All software is invoked on the command-line as follows: `perl program-name`. If arguments are required, a suitable error message is displayed.

(18) a. `Unigrams.pm`: collects unigram statistics from some textfile
b. `Bigrams.pm`: collects bigram statistics from some textfile
c. `uniapprox.pl`: creates some number of unigram approximations given some text
d. `biapprox.pl`: creates some number of bigram approximations given some text

## References

CHARNIAK, EUGENE. 1993. *Statistical Language Learning.* Cambridge: MIT Press.