

# 1 Considerations for the structure of lecture

1. I want to show you how you might develop a language model that suits a language-analysis problem you face.
2. The fewer hidden parameters in a model, the "easier" the problem of inferring those values from data.

# 2 Motivating example: modeling small-talk vs. non-small talk

Let's consider a generative story like the following: *(cf. ~~topic~~ topic modeling approach)*

1. Pick a sentence length  $l$ . *(easily generalized) - but why not "long" vs. "short" )*
2. Pick a sequence of  $l$  states: where the two possible state types are st for small talk, nst for not small-talk
3. For each state, pick a word according to that state's distribution over single words.

Example; we might decide we're going to say a five-word sentence, where the first word and the 4th and 5th words are going to be small-talk words.

## 2.1 Ideas for further refinement

- st might have a higher probability of ...
- st might have a higher probability of ...
- st might have a higher probability of

## 2.2 Sample data

Written "vertically" instead of "horizontally" to leave room to write on the sides.

Two sentences: *documents, sequences.*

hi  
i  
agree  
thanks  
bye

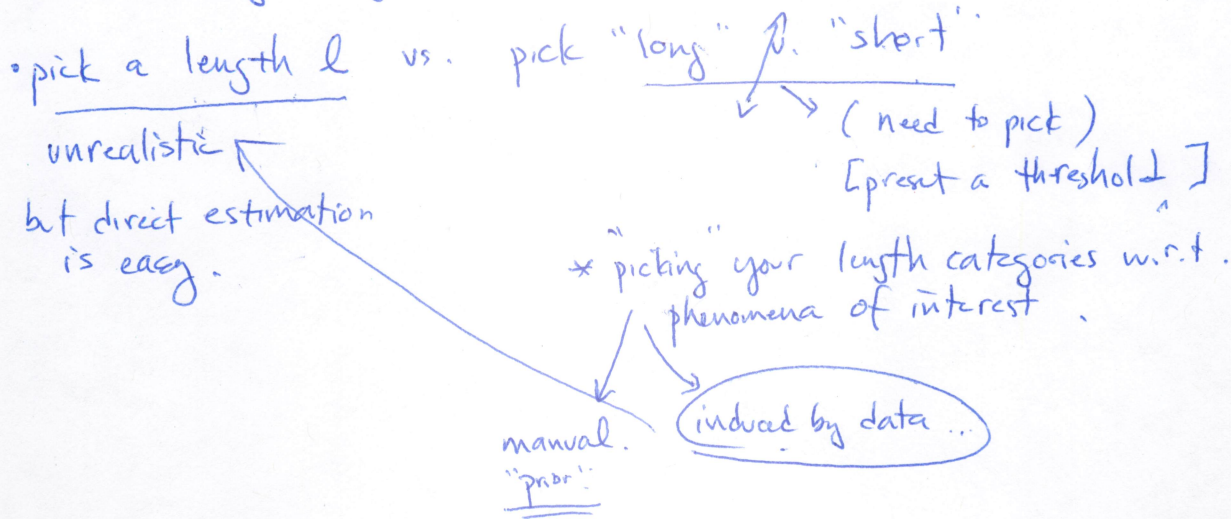
hi -  
sell  
hi [some stock ticker symbol]  
now  
thanks -

length & modeling... (§2, #1)

"middling"

lec 19  
11/9

PS 1



how come no other class worries about length modeling? (b/c no one has had one).  
q: moot.

example if you are not careful about modeling:

$w_1, w_2, \dots, w_n$   
↑  
a word.

$$p(w_1, w_2, \dots, w_n) = \prod_i p(w_i)$$

~~$p(\text{cats})$~~

$$\phi_{\text{cat}} = 1$$

$$P(\text{cat}) = 1$$

$$P(\text{catcat}) = 1 \times 1 = 1$$

these are not the same P's!

~~contradiction~~  
→ contradiction

what about priors? (length)

~~something~~:

< to be more accurate > - especially when sample seems limited or unrepresentative

• these are mathematically convenient priors:

ex: multinomial → Dirichlet prior

- interpolating: take as ~~prior~~ an LM built on generic English. Plug additional knowledge source

mine: ~~P<sub>st</sub>~~

free parameter  $\alpha \in [0, 1]$  Plug  

$$P(w) = \alpha P_{\text{st}}(w) + (1-\alpha) P_{\text{LM}}(w)$$

other ways to combine 2 LMs?

backoff: if you had an indicator that your special LM was good or not: ~~stg, stb~~

rely on  $P_{st}()$  when it's good

rely on  $P_{LMs}()$  when it's not good.

~~how~~ the details are in defining indicator, and making normalization work.

~~P(w)~~ ~~P(w)~~ ~~count~~ ~~data~~

ex: "switch" might be frequency of the word.

how do you evaluate, say,  $\alpha = .6$  vs.  $\alpha = .9$

- see which assigns more accurate probabilities

- you can often check: 500 words from ~~the~~ } held-out (not in inference data)  
500 words from ~~another~~ }

the data you want to model.

aside: don't compare probs of two samples of diff. lengths.

why?  $P(w_i) = \prod P_i(w)$

longer sentences usually less probable

from a ~~to~~ very diff. sample. reasonable.

$P_{\alpha=.6}$  (you want to model)

$P_{\alpha=.9}$  (you want to model)

$P_{\alpha=.6}$  (you don't want to model)

$P_{\alpha=.9}$  (out-of-domain data)