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:

1. Pick a sentence length $\ell$. (easily generalized) - but why not “long” vs. “short”?
2. Pick a sequence of $\ell$ 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:  
  
```
hi
i
agree
thanks
bye
```

```
hi  buy
sell
hi [some stock ticker symbol]
now
thanks
```
length vs modeling (§2, #1)  
- pick a length vs. pick "long" & "short"  
  unrealistic  
  but direct estimation is easy.

- "muddling"  
  need to pick  
  [present a threshold]  
  * picking your length categories w.r.t. phenomena of interest.
  manual "prior"  
  * induced by data
  
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:  
\[ p(w_1w_2...w_n) = \prod_i p(w_i) \]

- a word.

\[ p(cat) = 1 \]
\[ p(cat \_cat) = 1 \times 1 = 1 \]

\[ \phi_{cat} = 1 \]

\[ \phi \neq \phi \]
contradiction.

What about priors? (length)

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

- these are mathematically convenient priors & justified
  
  ex: multinomial \rightarrow Dirichlet prior
  
  - interpolating: take as \( P \) an LM built on generic English. 
    
    mine: \( P_{\text{est}} \) free parameter \( \alpha \in [0,1] \)
    
    \[ P(w) = \frac{\alpha P_{\text{LM}}(w) + (1-\alpha) P_{\text{est}}(w)}{\alpha P_{\text{LM}}(w) + (1-\alpha) P_{\text{est}}(w)} \]
other ways to combine 2 LMs?

backoff: if you had an indicator that your special LM was good or not:

relies on Pst( ) when it's good
relies on Pneg( ) when it's not good.

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

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

---

how do you evaluate, say, $\alpha = 0.6$ vs. $\alpha = 0.9$

- see which assigns more accurate probabilities
- you can often check: 500 words from hold-out (not in in inference set)

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

why? $P(w_e) = \frac{\sum_{i=1}^{T} p_i(w)}{T}$

longer sentences usually less probable

- $P\alpha = .6$ (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)