Introduction to language models

Roger Levy

9.19: Computational Psycholinguistics

Which did you hear?

Eyes awe of an

I saw a van

Which did you hear?

The sail of a boat

The sale of a boat

Which did you hear?

`It's not easy to wreck an ice beach`

`It's not easy to wreck a nice beach`

`It's not easy to recognize speech`

Which did you hear?

A dog's tale

A dog's tail

Shannon's guessing game

START

(Shannon, 1951; Taylor, 1953)

Radineg scralmbed wrods

in tehy All btahree. unooncuiscs stay be mmamals to to sttae for wehlas, need selep, buscaee long, they cnnaot an too conoscuis idnncilug but

All mmamals selep, idnncilug wehlas, but they cnnaot stay in an unooncuiscs sttae for too long, buscaee tehy need to be conoscuis to btahree.

Applications of language prediction

- In speech understanding, identify words incrementally!

cap tucked *captain*

- Especially challenging given ***segmentation ambiguity***

Robustness in comprehension

I uh, I found out that my grandmother was one of a twin.

I

a twin

a pair of twins

a set of twins

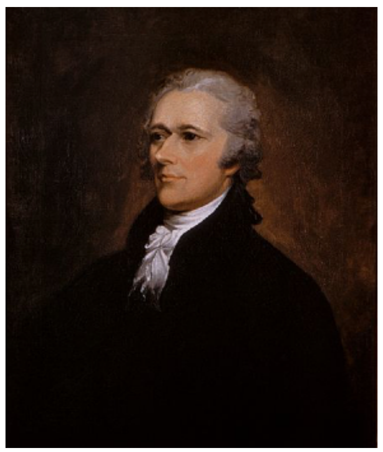
The businessman benefited the tax law significantly.

from

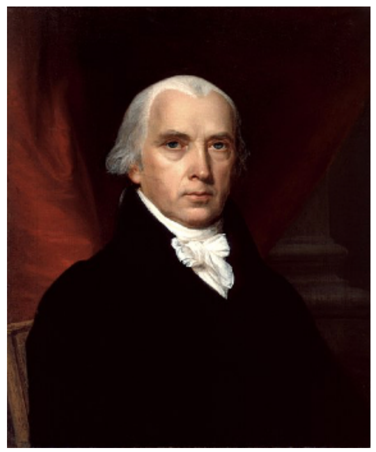
(parsed Switchboard corpus; Gibson et al., 2013)

Speaker modeling (e.g., author ID)

- One of the oldest applications of probability in computational linguistics!

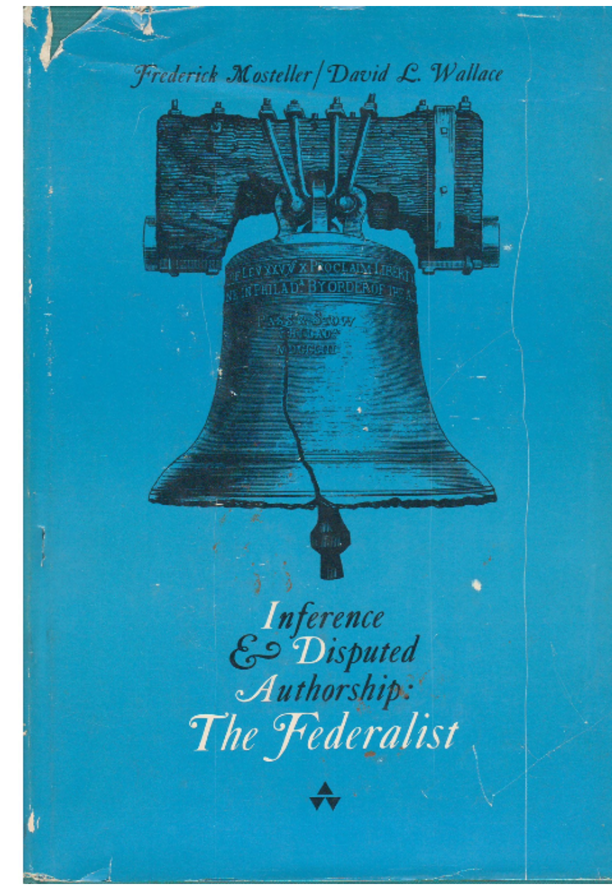


Alexander Hamilton



James Madison

John Jay



As the people are the only legitimate fountain of power, and it is from them that the constitutional charter, under which the several branches of government hold their power, is derived, it seems strictly consonant to the republican theory, to recur to the same original authority, not only whenever it may be necessary to enlarge, diminish, or new-model the powers of the government, but also whenever any one of the departments may commit encroachments on the chartered authorities of the others.
— Federalist 49, Publius

(Mosteller & Wallace, 1964)

Human comprehension difficulty

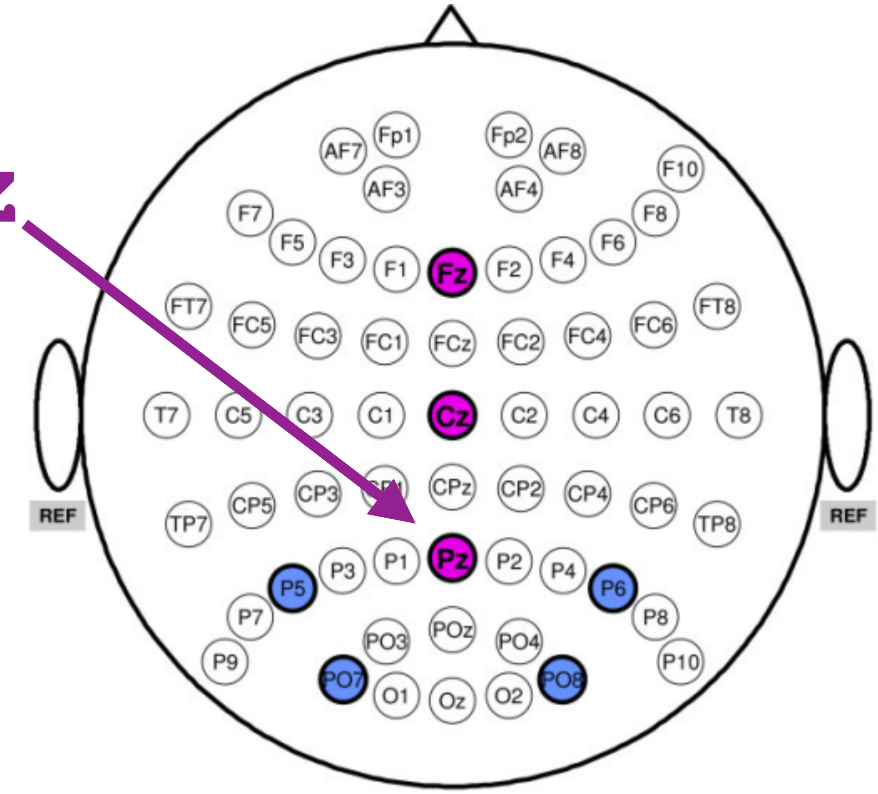
- Brains are ***prediction*** engines!

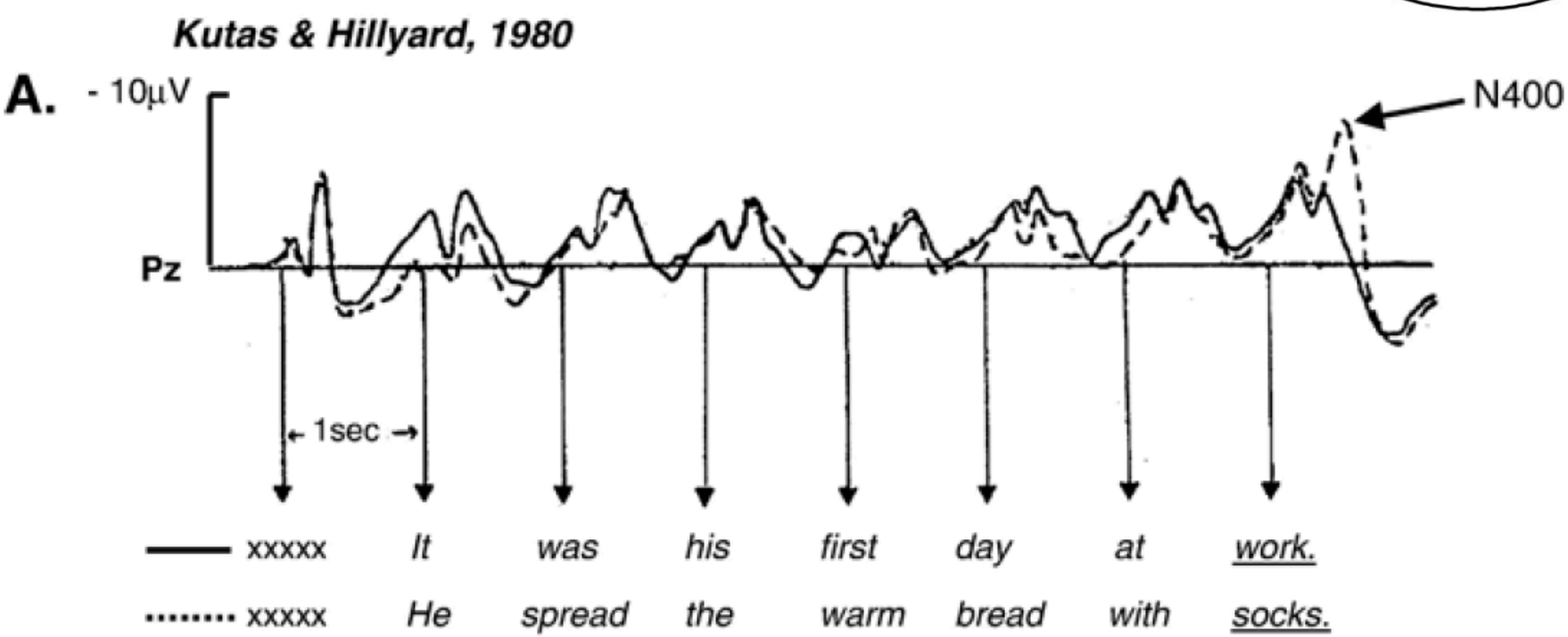
  my brother came inside to… *chat*? *wash*? *get warm*?

  the children went outside to… *play*

- Predictable words are read faster (Ehrlich & Rayner, 1981) and have distinctive EEG responses (Kutas & Hillyard 1980)
- The more we expect an event, the easier it is to process

Word responses





Encoding meaning into words

- Relevant for human language production, spoken dialog systems, machine translation, and more!

`dog's tail      6000:1   dog's tale`

`tail of a dog   750:1    tale of a dog`

Collocationality

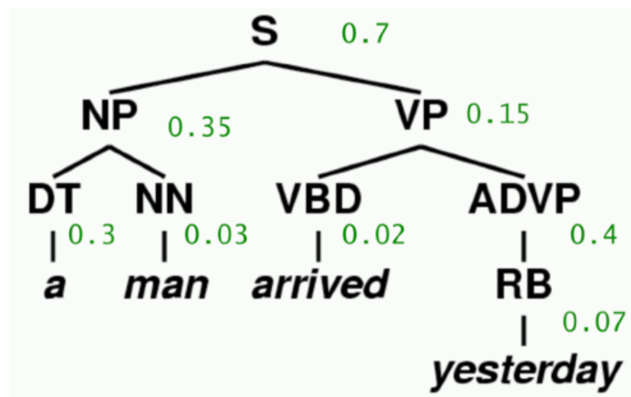
- A **collocation** is a word sequence that appears "unusually often"
- Consider the following word pairs in strength of the collocate:

`young childhood`	**`early childhood`**
`mass destruction`	**`illegal destruction`**
`good cuisine`	**`ethnic cuisine`**

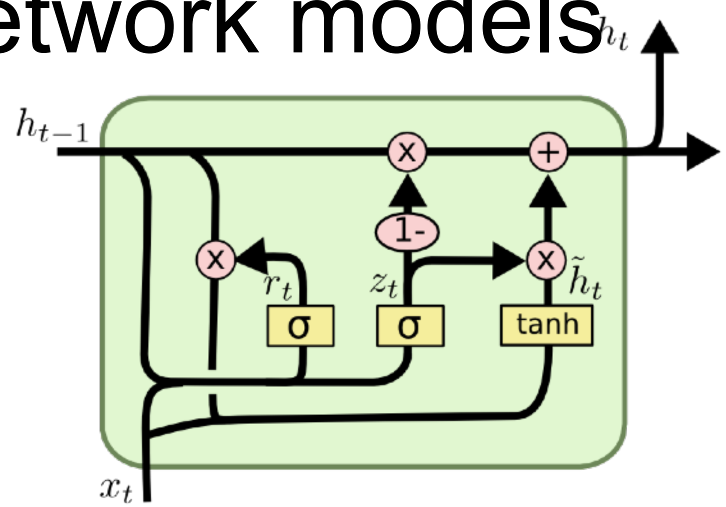
Word sequence frequencies

Modeling human knowledge of word sequences

- Many techniques, none perfect!
  - Probabilistic grammars



  - Neural network models



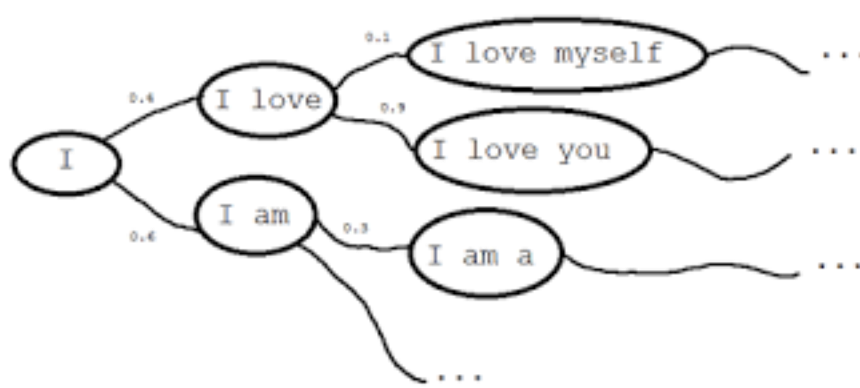
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

  - *n*-gram models



Today

n-grams from chain rule decomposition

- Probability that next sentence is "dogs chase cats"?

$$P(\vec{w} = \texttt{\$ dogs chase cats \$})$$

- Remember the chain rule!

$$P(x_1, \ldots, x_k) = \prod_{i=1}^{k} P(x_i | x_1, \ldots, x_{i-1})$$

- Applying this to our sentence we get

$$\begin{aligned} P(\vec{w} = \texttt{\$ dogs chase cats \$}) = & P(\texttt{\$}|\texttt{\$ dogs chase cats})\times \\ & P(\texttt{cats}|\texttt{\$ dogs chase})\times \\ & P(\texttt{chase}|\texttt{\$ dogs})\times \\ & P(\texttt{dogs}|\texttt{\$}) \end{aligned}$$

- Simplify—e.g., assume $w_i \perp w_{1\ldots i-2} | w_{i-1}$ to give us

$$P(\texttt{\$ dogs chase cats \$}) \approx P(\texttt{\$}|\texttt{cats})P(\texttt{cats}|\texttt{chase})P(\texttt{chase}|\texttt{dogs})P(\texttt{dogs}|\texttt{\$})$$

- MARKOV ASSUMPTION, giving a **2**-gram (**bi**gram) model

n-gram approximations of Shakespeare

1 gram	–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have –Hill he late speaks; or! a more to leg less first you enter
2 gram	–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. –What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done. –This shall forbid it should be branded, if renown made it empty.
4 gram	–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in; –It cannot be but so.

(courtesy Dan Jurafsky)

n-gram approximations of the Wall Street Journal

1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	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
3 gram	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

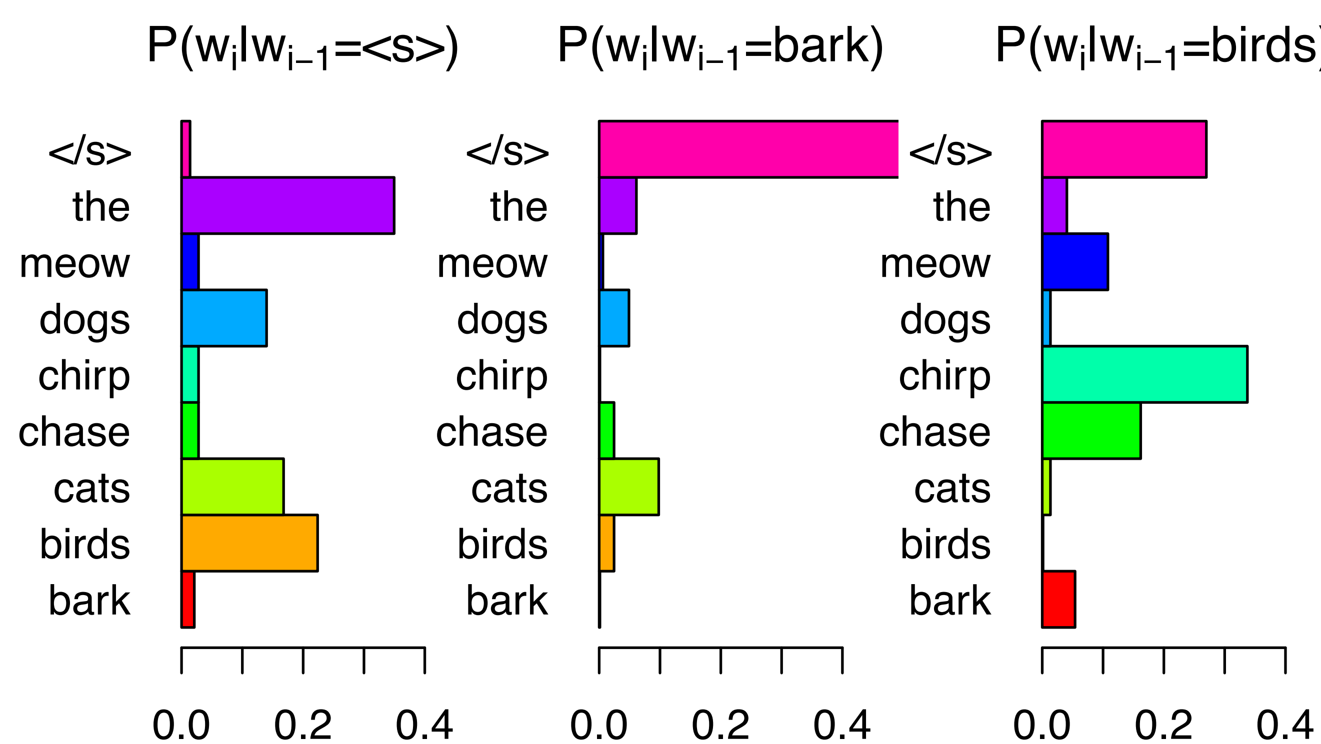
(courtesy Dan Jurafsky)

Maximum likelihood n-gram estimation

- **General scenario:**
  - You want to estimate conditional probabilities $P(Y|X)$
  - You have training data consisting of some $\langle X, Y \rangle$-pairs
  - You have chosen a "model class" (a PARAMETERIZED FAMILY of probability distributions)

- **Bigram estimation:**
  - You want to estimate $P(w_i|w_{i-1})$ in a language model
  - You have some sentences
  - You assume each w_{i-1} has its own multinomial over w_i

```
<s> dogs chase cats </s>
<s> dogs bark </s>
<s> cats meow </s>
<s> dogs chase birds </s>
<s> cats chase birds </s>
<s> dogs chase the cats </s>
<s> the birds chirp </s>
```



(repeat slide from lecture 3)

and so forth…

Maximum likelihood estimation

```
<s> dogs chase cats </s>
<s> dogs bark </s>
<s> cats meow </s>
<s> dogs chase birds </s>
<s> cats chase birds </s>
<s> dogs chase the cats </s>
<s> the birds chirp </s>
```

$c(w_{i-1}=\text{dogs}, w_i=\text{chase})$	= 3
$c(w_{i-1}=\text{dogs}, w_i=\text{bark})$	= 1
$c(w_{i-1}=\text{dogs})$	= 4

- Consider each multinomial parameter
  - e.g., let us call p the value of $P(w_i=\text{bark}|w_{i-1}=\text{dogs})$
  - So the value of $P(w_i \neq \text{bark}|w_{i-1}=\text{dogs})$ is $1-p$
  - Likelihood for the part of the data where w_{i-1}=dogs:

```
wi-1 wi
dogs chase
dogs bark
dogs chase
dogs chase
```

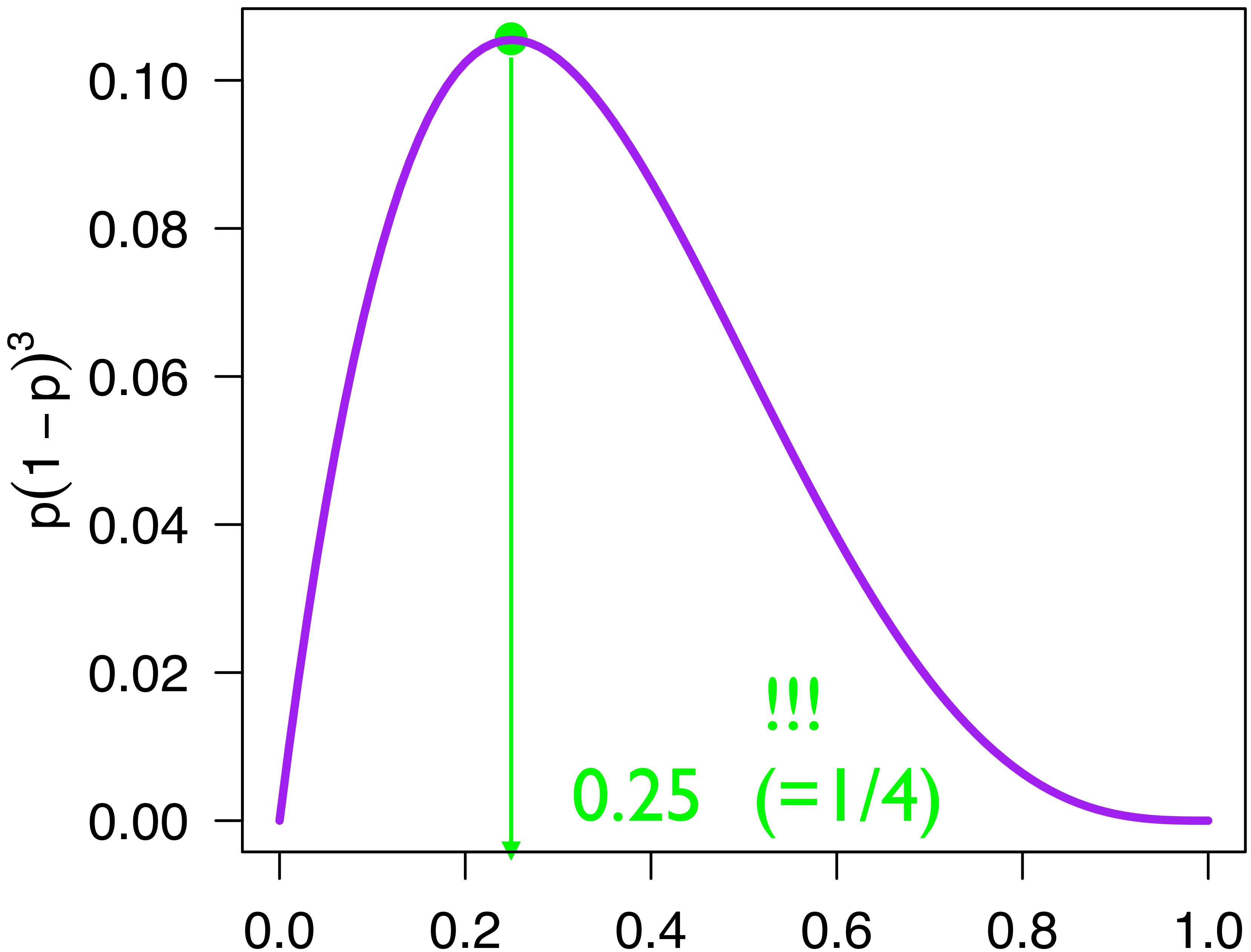
$$p(1-p)^3$$

(repeat slide from lecture 3)

Maximum likelihood estimation

- *p* refers to the value of P(w_i=bark|w_{i-1}=dogs)
- Likelihood for that part of data where w_{i-1}=dogs:

w_{i-1}	w_i
dogs	chase
dogs	bark
dogs	chase
dogs	chase



This is choosing the *maximum likelihood estimate* (**MLE**)

The **MLE** also turns out to be the *relative frequency estimate* (RFE)

(repeat slide from lecture 3)

Why smooth *n*-gram models?

Training data (bigram-counts representation):

```
Context the, events:  cats:  1 birds:  1
Context meow, events:  </s>:  1
Context birds, events:  chirp:  1 </s>:  2
Context chirp, events:  </s>:  1
Context cats, events:  meow:  1 </s>:  2 chase:  1
Context bark, events:  </s>:  1
Context </s>, events:  the:  1 cats:  2 dogs:  4
Context dogs, events:  bark:  1 chase:  3
Context chase, events:  the:  1 cats:  1 birds:  2
```

Held-out data:

`</s> birds chirp </s>` ← *unseen bigram*

Maximum-likelihood estimation gives *no* generalization to unseen events in the *n*-gram representation

Idea 1: additive smoothing

- Add a "pseudo"-count to each <context,event> pair

$$\widehat{P}_{\text{Laplace}}(w_i | w_{i-n+1} \ldots w_{i-1}) = \frac{\text{Count}(w_{i-n+1} \ldots w_{i-1} w_i) + 1}{\text{Count}(w_{i-n+1} \ldots w_{i-1}) + V}$$

vocabulary size

w_{-1}	w_i	Count
dogs	</s>	0
dogs	bark	1
dogs	birds	0
dogs	chase	3
dogs	dogs	0
dogs	the	0

$\widehat{P}_{MLE}(\text{</s>}|\text{bark}) = 1$

$\widehat{P}_{Laplace}(\text{</s>}|\text{bark}) = \frac{1}{6}$

- Too much added probability mass for rare (i.e., typical) contexts!

Generalized additive smoothing

- We can also add less than 1 to each count

$$\widehat{P}_{\text{Laplace}}(w_i|w_{i-n+1}\dots w_{i-1}) = \frac{\text{Count}(w_{i-n+1}\dots w_{i-1}w_i) + \lambda}{\text{Count}(w_{i-n+1}\dots w_{i-1}) + \lambda V}$$

- But this doesn't turn out to do so great in practice, either (we'll see in practicum)
- Fundamental issue: we should make different generalizations about:
  - different contexts;
  - and different events.
- Additive smoothing accomplishes neither of these

Idea 2: model interpolation

- Suppose we have a **unigram model** and we also have a **bigram model**
- We could mix the two models' probabilities together:

$$P_{\text{Interpolated}}(w_i|w_{i-1}) = \lambda P(w_i|w_{i-1}) + (1-\lambda)P(w_i)$$

- This modification of a standard bigram model **makes different generalizations about different events**
  - How?
- Words that are more frequent overall become more expected regardless of context
- Interpolation weights can also be a function of context:

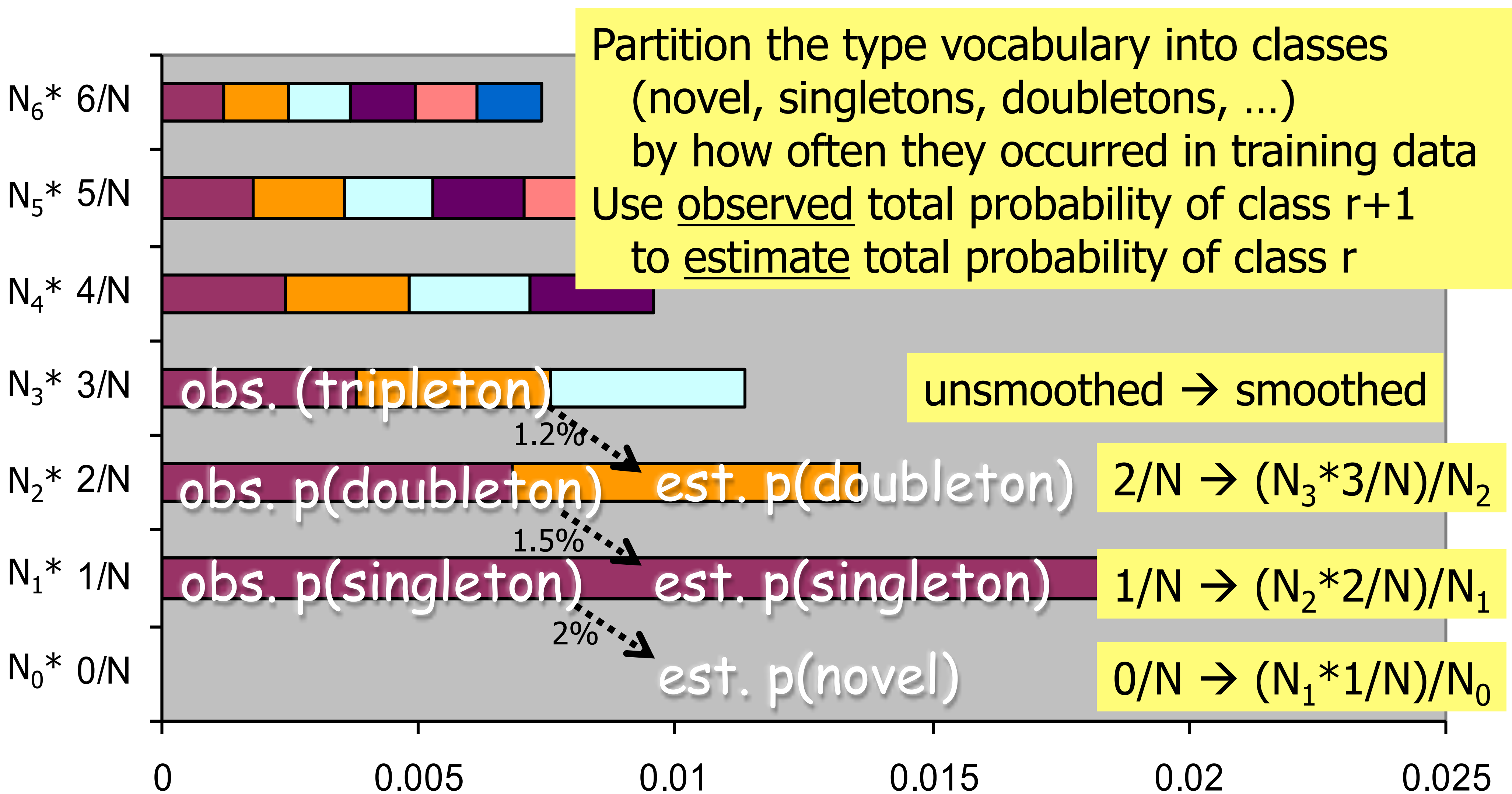
$$P_{\text{Interpolated}}(w_i|w_{i-1}) = \lambda(w_{i-1})P(w_i|w_{i-1}) + (1-\lambda(w_{i-1}))P(w_i)$$

- And we can extend this approach to higher-order *n*-grams

Idea 3: Leveraging a context's type diversity

- The more rare events a context has, the more new events we should expect!

Good-Turing Smoothing Idea



(Courtesy Jason Eisner)

$r/N = (N_r * r/N)/N_r \rightarrow (N_{r+1} * (r+1)/N)/N_r$

Idea 4: leveraging an event's context diversity

*I can't see without my reading*___________ Francisco glasses

Define the **continuation probability** of a word as the number of <context,word> pairs it completes

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\left|\{(w_{j-1}, w_j) : c(w_{j-1}, w_j) > 0\}\right|}$$

(example courtesy Dan Jurafsky)

Kneser-Ney smoothing

$$P_{KN}(w_i \mid w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{CONTINUATION}(w_i)$$

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})} \left| \{ w : c(w_{i-1}, w) > 0 \} \right|$$

Ideas we haven't implemented yet

- **Generalizing across contexts or events** in terms of their similarity to one another
- **Varying the window of context** that we consider
- Representing "**proximity**" to the event **in non-linear terms**