Natural Language Processing

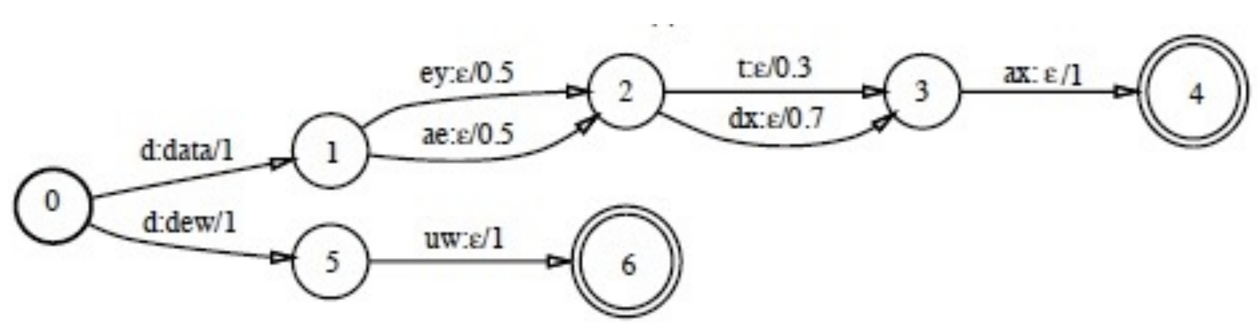
Spring 2017

Unit 1: Sequence Models

Lectures 5-6: Language Models and Smoothing

required

optional

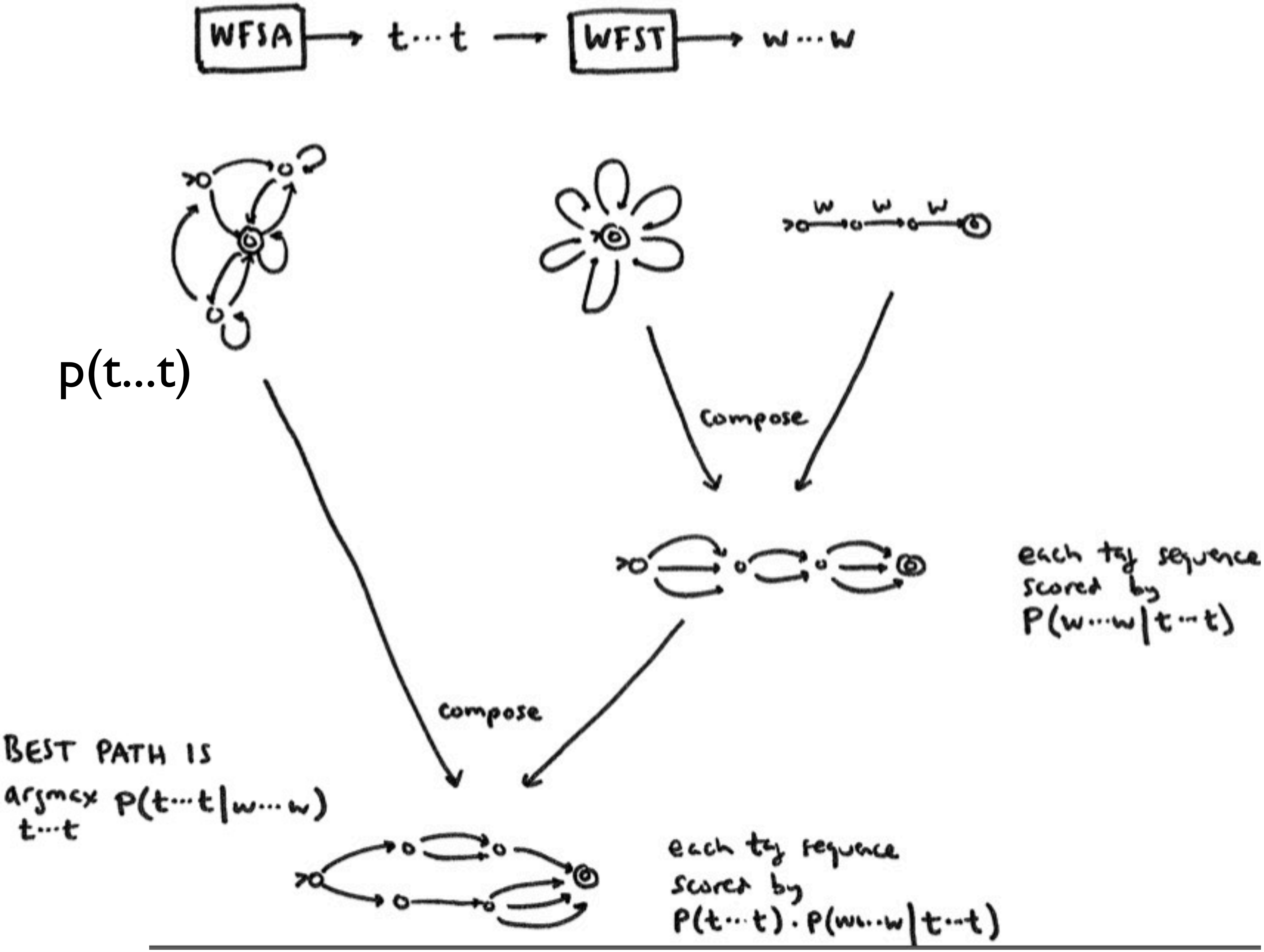


Professor Liang Huang

liang.huang.sh@gmail.com

Noisy-Channel Model

Noisy-Channel Model



p(t...t)

Applications of Noisy-Channel

Application	Input	Output	p(i)	p(o|i)
Machine Translation	L_1 word sequences	L_2 word sequences	p(L_1) in a language model	translation model
Optical Character Recognition (OCR)	actual text	text with mistakes	prob of language text	model of OCR errors
Part Of Speech (POS) tagging	POS tag sequences	English words	prob of POS sequences	p(w|t)
Speech recognition	word sequences	speech signal	prob of word sequences	acoustic model
spelling correction	correct text	text with mistakes	prob. of language text	noisy spelling

Noisy Channel Examples

Th qck brwn fx jmps vr th lzy dg.
Ths sntnc hs ll twnty sx lttrs n th lphbt.

I cnduo't bvleiee taht I culod aulaclty uesdtannrd waht I was rdnaieg. Unisg the icndeblire pweor of the hmuan mnid, aocdcrnig to rseecrah at Cmabrigde Uinervtisy, it dseno't mttaer in waht oderr the lterets in a wrod are, the olny irpoamtnt tihng is taht the frsit and lsat ltteer be in the rhgit pclae.

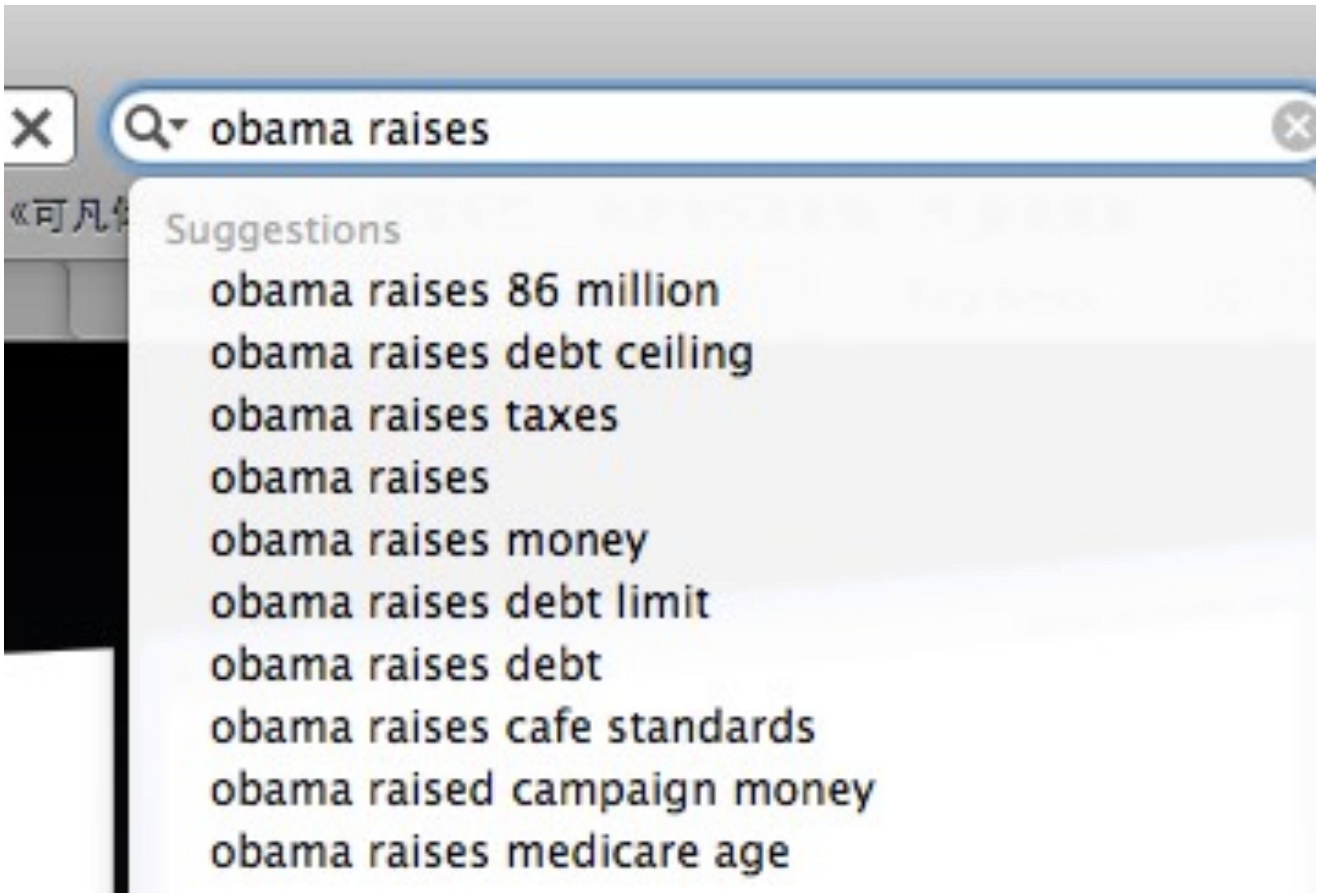
Therestcanbeatotalmessandyoucanstillreaditwithoutaproblem.Thisisbecausethehumanminddoesnotreadeveryletterbyitself,butthewordasawhole.

研表究明，汉字的序顺并不定一能影阅响读，比如当你看完句这话后，才发这现里的字全是都乱的。

研究表明，汉字的顺序并不一定能影响阅读，比如当你看完这句话后，才发现这里的字全都是乱的。

Language Model for Generation

- search suggestions



Language Models

- problem: what is $P(\mathbf{w}) = P(w_1 \, w_2 \ldots w_n)$?
  - ranking: P(an apple) > P(a apple)=0, P(he often swim)=0
  - prediction: what's the next word? use $P(w_n \mid w_1 \ldots w_{n-1})$
    - Obama gave a speech about ____ .

sequence prob, not just joint prob.

- $P(w_1 \, w_2 \ldots w_n) = P(w_1) \, P(w_2 \mid w_1) \ldots P(w_n \mid w_1 \ldots w_{n-1})$
- $\approx P(w_1) \, P(w_2 \mid w_1) \, P(w_3 \mid w_1 \, w_2) \ldots P(w_n \mid w_{n-2} \, w_{n-1})$ trigram
- $\approx P(w_1) \, P(w_2 \mid w_1) \, P(w_3 \mid w_2) \ldots P(w_n \mid w_{n-1})$ bigram
- $\approx P(w_1) \, P(w_2) \, P(w_3) \ldots P(w_n)$ unigram
- $\approx P(w) \, P(w) \, P(w) \ldots P(w)$ 0-gram

Estimating *n*-gram Models

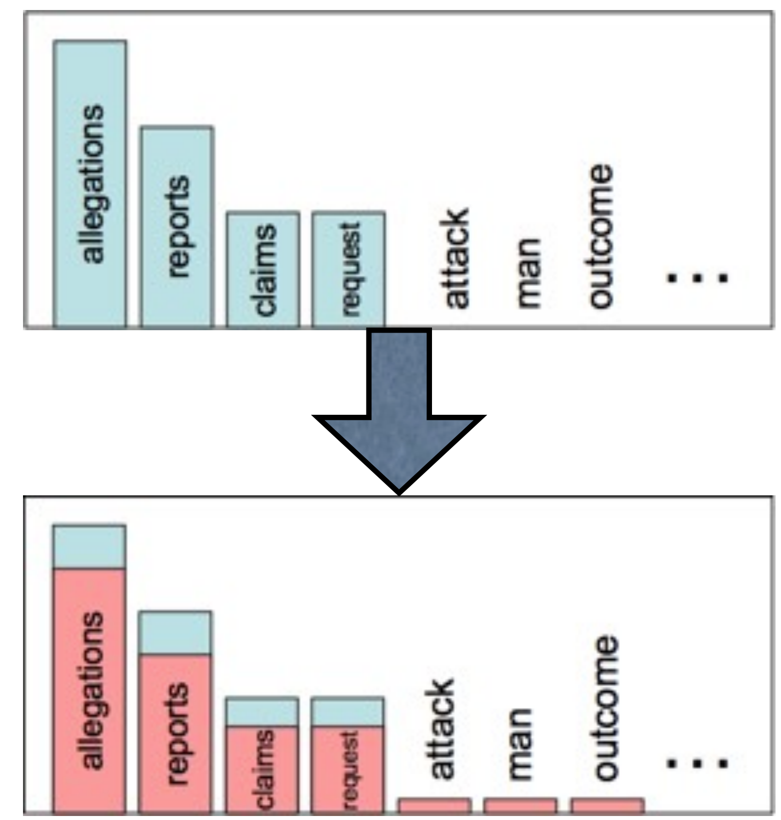
	" In	person	she	was	inferior ~~Superior~~	to	both	sisters "		
0-gram			10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	≈	10^{-36}
unigram			.011	.015	.00005	.032	.0005	.0003	=	4×10^{-17}
bigram			.009	.122	0	.212	.0004	.006	=	
trigram			?	.5	0	?	0	0	=	
4-gram			?	?	0	?	?	?	=	

(textbook, table 6.3)

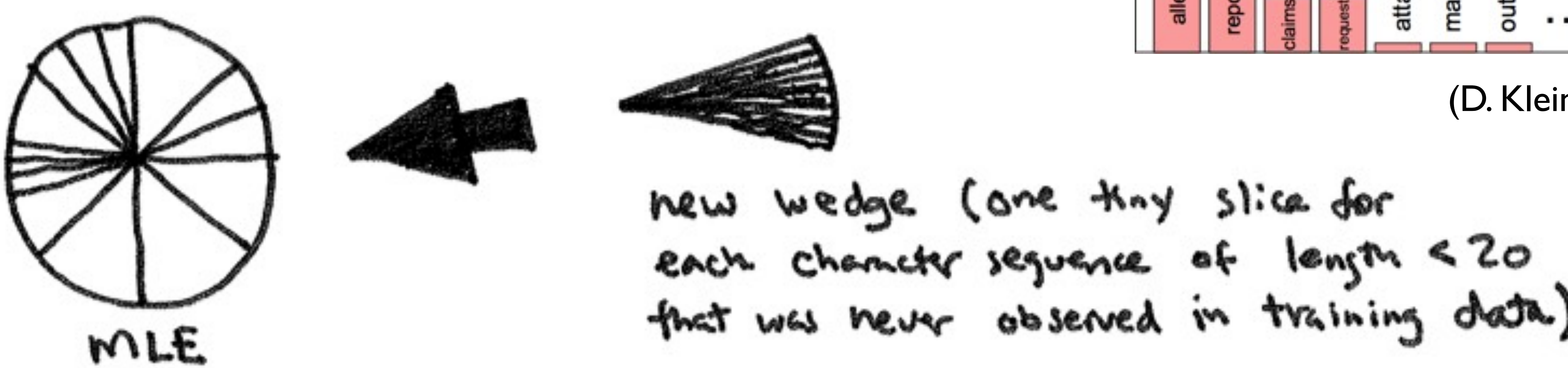
- maximum likelihood: $p_{ML}(x) = c(x)/N$; $p_{ML}(xy) = c(xy)/c(x)$
- problem: unknown words/sequences (unobserved events)
- sparse data problem
- solution: smoothing

Smoothing

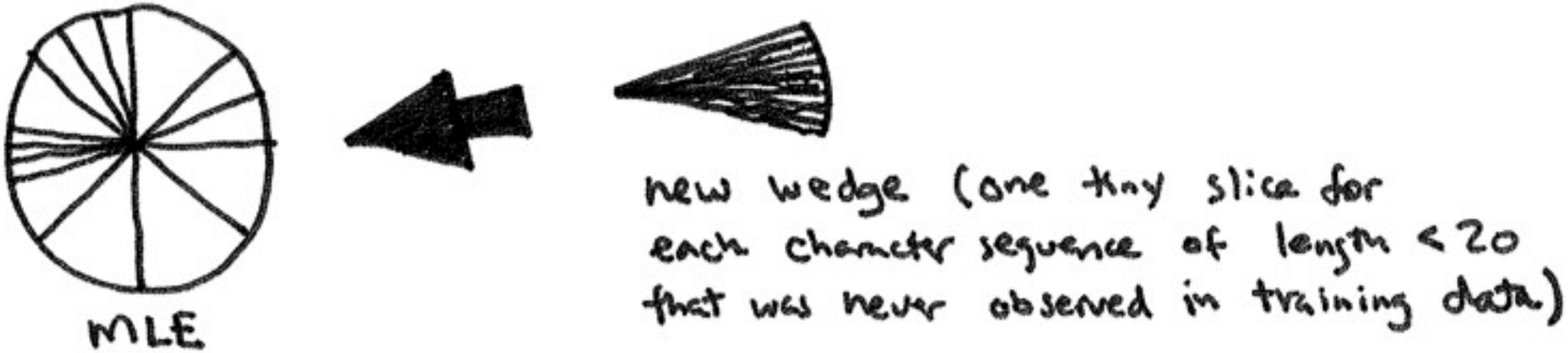
- have to give some probability mass to unseen events
  - (by discounting from seen events)
- Q1: how to divide this wedge up?
- Q2: how to squeeze it into the pie?



(D. Klein)



Smoothing: Add One (Laplace)



- MAP: add a "pseudocount" of 1 to every word in Vocab
- $P_{lap}(x) = (c(x) + 1) / (N + V)$ V is Vocabulary size
  - $P_{lap}(unk) = 1 / (N+V)$ same probability for all unks
- how much prob. mass for unks in the above diagram?
  - e.g., $N=10^6$ tokens, $V=26^{20}$, $V_{obs} = 10^5$, $V_{unk} = 26^{20} - 10^5$

Smoothing: Add Less than One

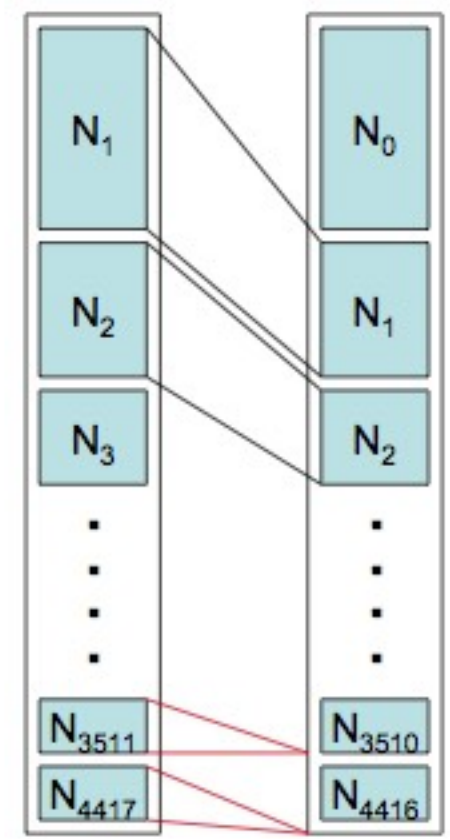
- add one gives too much weight on unseen words!
- solution: add less than one (Lidstone) to each word in V
- $P_{lid}(x) = (c(x) + \lambda) / (N + \lambda V)$ $0<\lambda<1$ is a parameter
  - $P_{lid}(unk) = \lambda / (N+ \lambda V)$ still same for unks, but smaller
- Q: how to tune this λ ? on held-out data!

Smoothing: Witten-Bell

- key idea: use one-count things to guess for zero-counts
  - recurring idea for unknown events, also for Good-Turing
- prob. mass for unseen: T / (N + T) T: # of seen types
  - 2 kinds of events: one for each token, one for each type
  - = MLE of seeing a new type (T among N+T are new)
  - divide this mass evenly among V-T unknown words
- $p_{wb}(x) = T / (V-T)(N+T)$ unknown word
  $= c(x) / (N+T)$ known word
- bigram case more involved; see J&M Chapter for details

Smoothing: Good-Turing

- again, one-count words in training ~ unseen in test
- let N_c = # of words with frequency r in training
- $P_{GT}(x) = c'(x) / N$ where $c'(x) = (c(x)+1) N_{c(x)+1} / N_{c(x)}$
- total adjusted mass is $\text{sum}_c\ c'\ N_c = \text{sum}_c\ (c+1)\ N_{c+1} / N$
  - remaining mass: N_1 / N: split evenly among unks



EXAMPLE:

r	Nr	Nr+1	r*	r*/N
0	1000	100	-	1-Z
1	100	40	0.8	Sums to Z
2	40	20	1.5	
3	20	10	2.0	
4	10	6	3.0	
5	6	3	3.0	
⋮	⋮	⋮	⋮	⋮

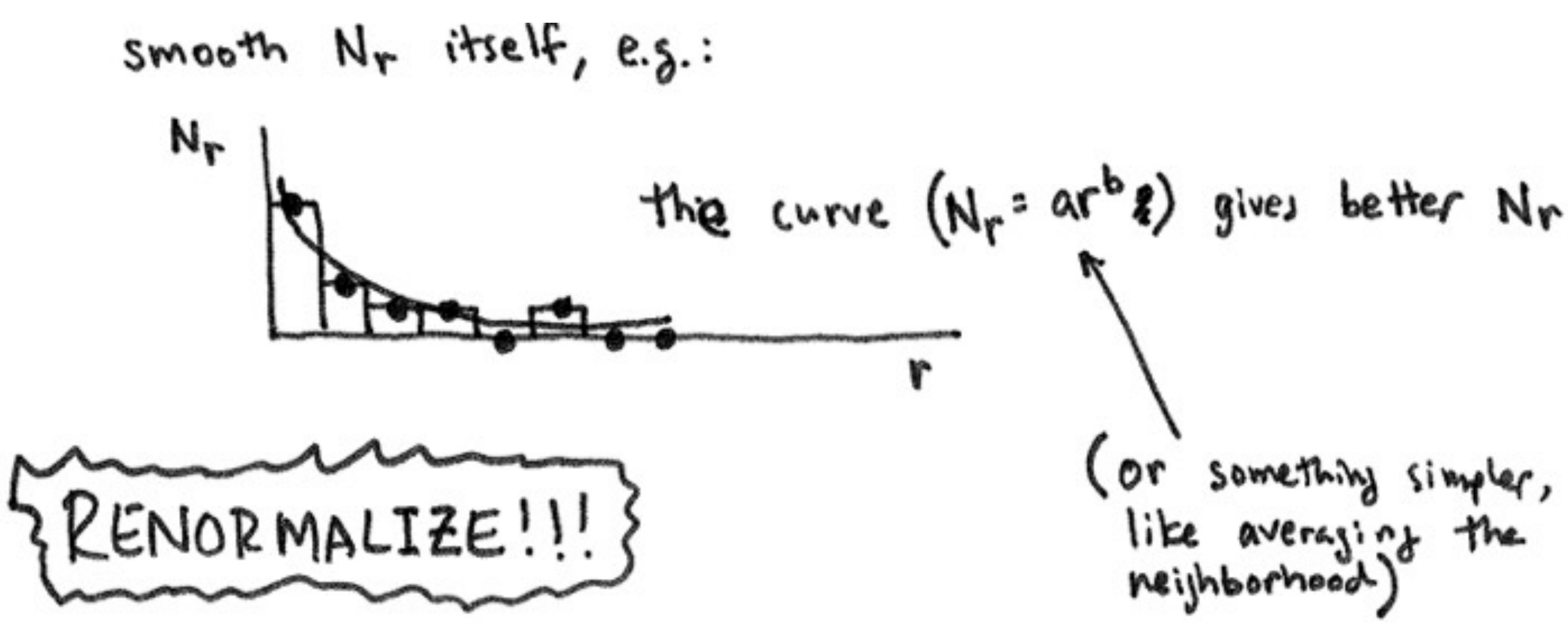
Smoothing: Good-Turing

- from Church and Gale (1991).
  bigram LMs.   unigram vocab size = 4×10^5.
  T_r is the frequencies in the held-out data (see $f_{empirical}$).

$r = f_{MLE}$	$f_{empirical}$	f_{Lap}	f_{del}	f_{GT}	N_r	T_r
0	0.000027	0.000137	0.000037	0.000027	74 671 100 000	2 019 187
1	0.448	0.000274	0.396	0.446	2 018 046	903 206
2	1.25	0.000411	1.24	1.26	449 721	564 153
3	2.24	0.000548	2.23	2.24	188 933	424 015
4	3.23	0.000685	3.22	3.24	105 668	341 099
5	4.21	0.000822	4.22	4.22	68 379	287 776
6	5.23	0.000959	5.20	5.19	48 190	251 951
7	6.21	0.00109	6.21	6.21	35 709	221 693
8	7.21	0.00123	7.18	7.24	27 710	199 779
9	8.26	0.00137	8.18	8.25	22 280	183 971

Smoothing: Good-Turing

- Good-Turing is much better than add (less than) one
- problem 1: $N_{cmax+1} = 0$, so $c'_{max} = 0$
  - solution: only adjust counts for those less than k (e.g., 5)
- problem 2: what if $N_c = 0$ for some middle c?
  - solution: smooth N_c itself



Smoothing: Backoff & Interpolation

$$\hat{p}(w_i|w_{i-2}w_{i-1}) = \begin{cases} \tilde{p}(w_i|w_{i-2}w_{i-1}), & \text{if} C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha_1 p(w_i|w_{i-1}), & \text{if} C(w_{i-2}w_{i-1}w_i) = 0 \\ & \text{and} C(w_{i-1}w_i) > 0 \\ \alpha_2 p(w_i), & \text{otherwise.} \end{cases}$$

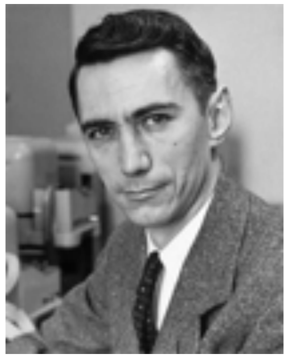
$$\begin{aligned} \hat{p}(w_i|w_{i-2}w_{i-1}) &= \lambda_1 p(w_i|w_{i-2}w_{i-1}) \\ &+ \lambda_2 p(w_i|w_{i-1}) \\ &+ \lambda_3 p(w_i) \end{aligned}$$

subject to the constraint that $\sum_j \lambda_j = 1$

Entropy and Perplexity

- classical entropy (uncertainty): $H(X) = -\text{sum } p(x) \log p(x)$
  - how many "bits" (on average) for encoding
- sequence entropy (distribution over sequences):
  - $H(L) = \lim 1/n \, H(w_1 \ldots w_n)$ (for language L) Q: why 1/n?
  - $= \lim 1/n \, \text{sum}\_\{w \text{ in } L\} \, p(w_1 \ldots w_n) \log p(w_1 \ldots w_n)$
- Shannon-McMillan-Breiman theorem:
  - $H(L) = \lim -1/n \log p(w_1 \ldots w_n)$ no need to enumerate w in L!
  - if w is long enough, just take $-1/n \log p(w)$ is enough!
- perplexity is $2^{H(L)}$

Entropy/Perplexity of English



- on 1.5 million WSJ test set:
  - unigram: 962 9.9 bits
  - bigram: 170 7.4 bits
  - trigram: 109 6.8 bits
- higher-order n-grams generally has lower perplexity
  - but hitting diminishing returns after n=5
  - even higher order: data sparsity will be a problem!
  - recurrent neural network (RNN) LM will be better
- what about human??

Shannon Papers

- Shannon, C. E. (1938). A Symbolic Analysis of Relay and Switching Circuits. *Trans. AIEE.* 57 (12): 713–723. cited ~1,200 times. (*MIT MS thesis*)
- Shannon, C. E. (1940). An Algebra for Theoretical Genetics. *MIT PhD Thesis*. cited 39 times.
- Shannon, C.E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, 1948. cited ~100,000 times.
- Shannon, C.E. (1951). Prediction and Entropy of Printed English. (*same journal*)
  - http://languagelog.ldc.upenn.edu/myl/Shannon1950.pdf cited ~2,600 times.

	F_0	F_1	F_2	F_3	F_{word}
26 letter	4.70	4.14	3.56	3.3	2.62
27 letter	4.76	4.03	3.32	3.1	2.14

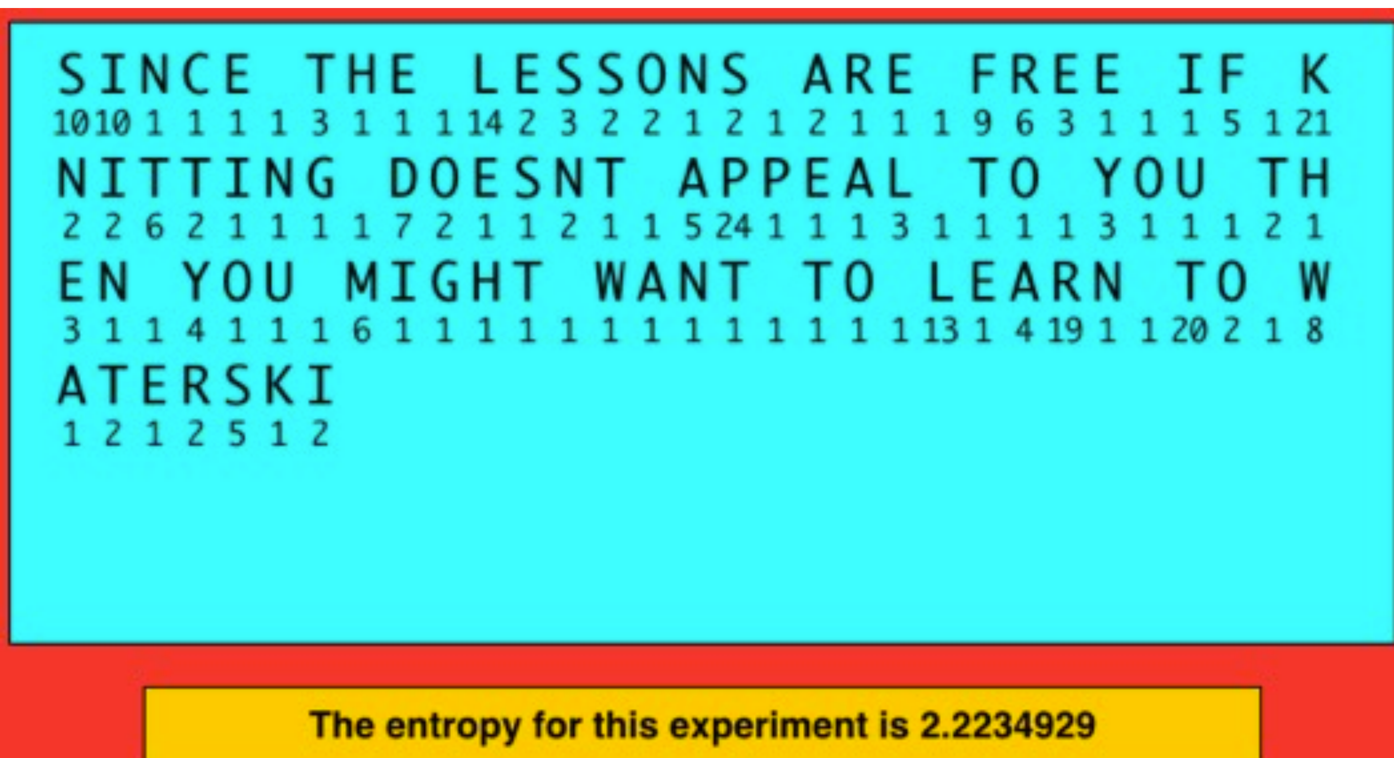
Zero-order approximation	XFOML RXKHRJFFJUJ ALPWXFWJXYJ FFJEYVJCQSGHYD QPAAMKBZAACIBZLKJQD
First-order approximation	OCRO HLO RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
Second-order approximation	ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE
Third-order approximation	IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

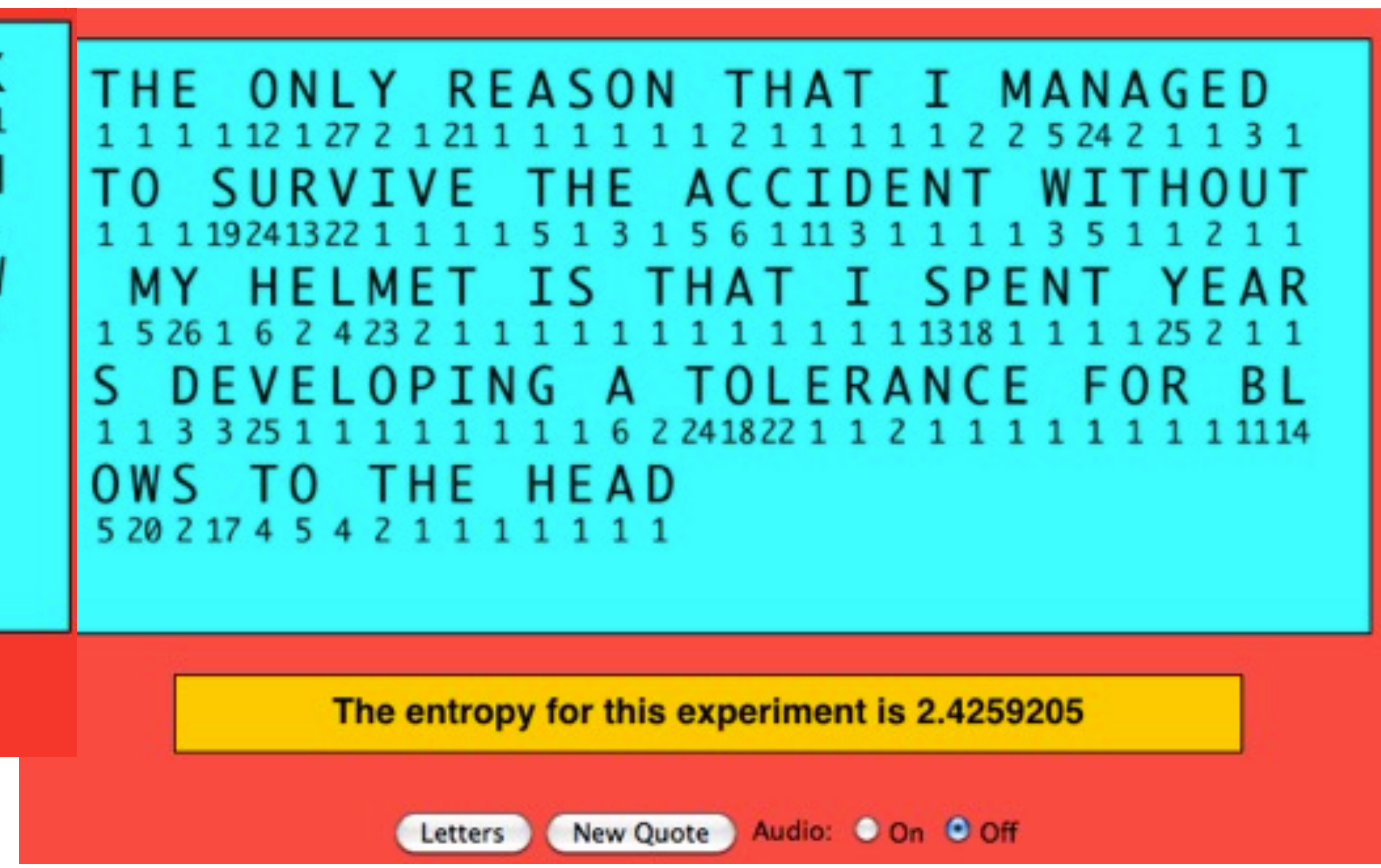
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

http://people.seas.harvard.edu/~jones/cscie129/papers/stanford_info_paper/entropy_of_english_9.htm

Shannon Game

- guess the next letter; compute entropy (bits per char)
- 0-gram: 4.76,   1-gram: 4.03,   2-gram: 3.32,   3-gram: 3.1
- native speaker: ~1.1 (0.6~1.3);   me: upperbound ~2.3





Q: formula for entropy?
(only computes upperbound)

http://math.ucsd.edu/~crypto/java/ENTROPY/

From Shannon Game to Entropy

T	h	e	_	b	r	o	k	e	n	_	v
2	1	1	1	11	3	2	5	1	1	1	15

The subject's identical twin would be able to reconstruct the original text from the guess sequence, so in that sense, it contains the same amount of information.

Let $c_1, c_2, \ldots c_n$ represent the character sequence, let $g_1, g_2, \ldots g_n$ represent the guess sequence, and let j range over guess numbers from 1 to 95, the number of printable English characters plus newline. Shannon [3] provides two results.

(*Upper Bound*). The entropy of $c_1, c_2, \ldots c_n$ is no greater than the unigram entropy of the guess sequence:

$$-\frac{1}{n}\log(\prod_{i=1} P(g_i)) = -\frac{1}{n}\sum_{i=1}^{n}\log(P(g_i)) = -\sum_{j=1}^{95} P(j)\log(P(j))$$

This is because this unigram entropy is an upper bound on the entropy of $g_1, g_2, \ldots g_n$, which equals the entropy of $c_1, c_2, \ldots c_n$. In human experiments, Shannon obtains an upper bound of 1.3 bits per character (bpc) for English, significantly better than the character n-gram models of his time (e.g., 3.3 bpc for trigram).

(*Lower Bound*). The entropy of $c_1, c_2, \ldots c_n$ is no less than:

$$\sum_{j=1}^{95} j \cdot [P(j) - P(j+1)] \cdot \log(j)$$

with the proof given in his paper. Shannon reported a lower bound of 0.6 bp

$$\sum_{i=1}^{27} i(q_i^N - q_{i+1}^N)\log i \leq F_N \leq -\sum_{i=1}^{27} q_i^N \log q_i^N. \qquad (17)$$

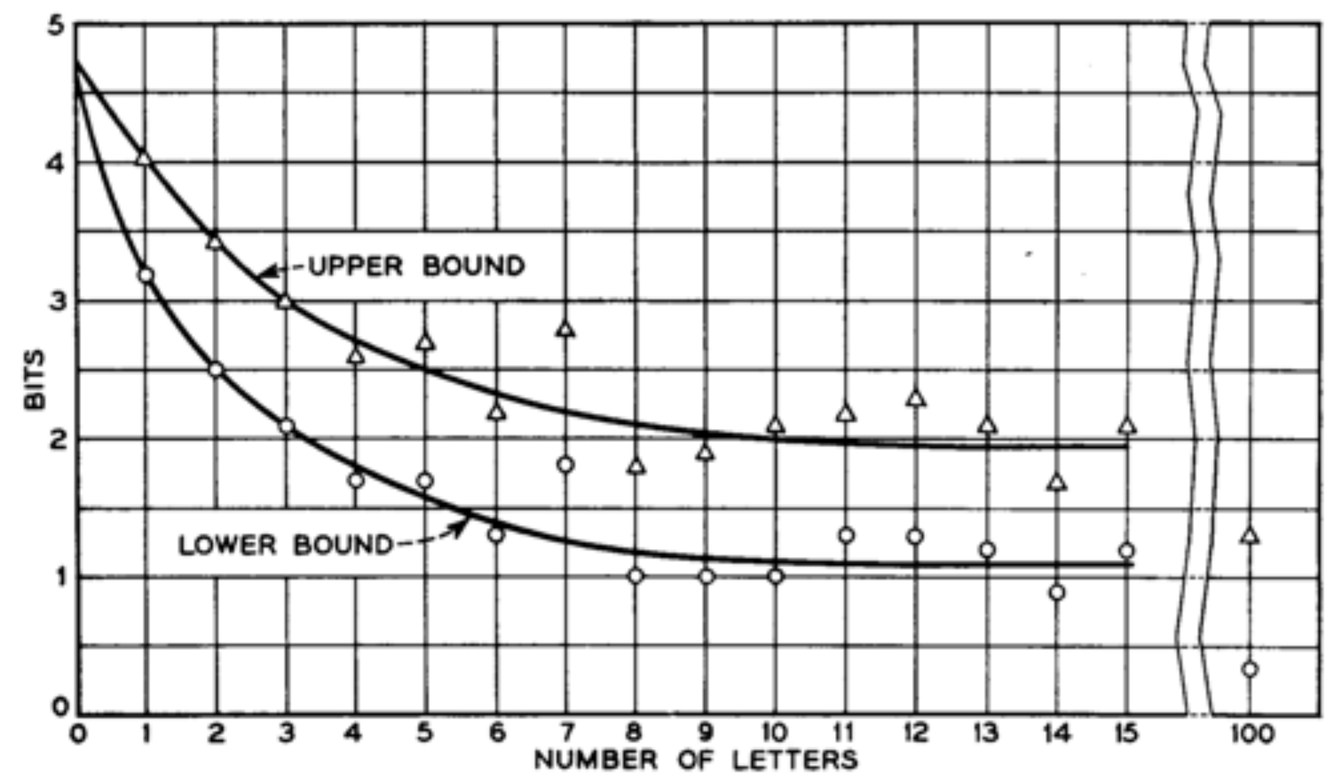
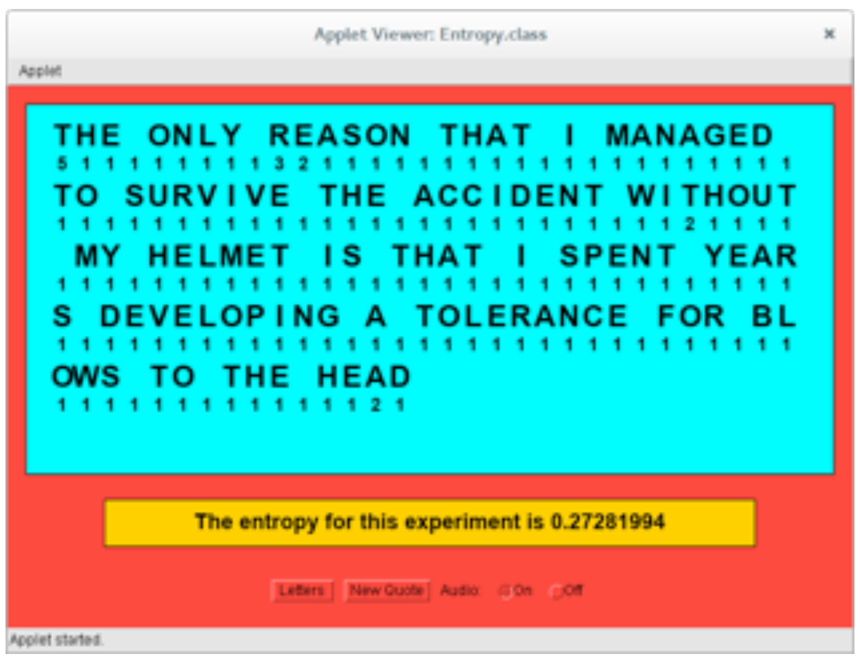
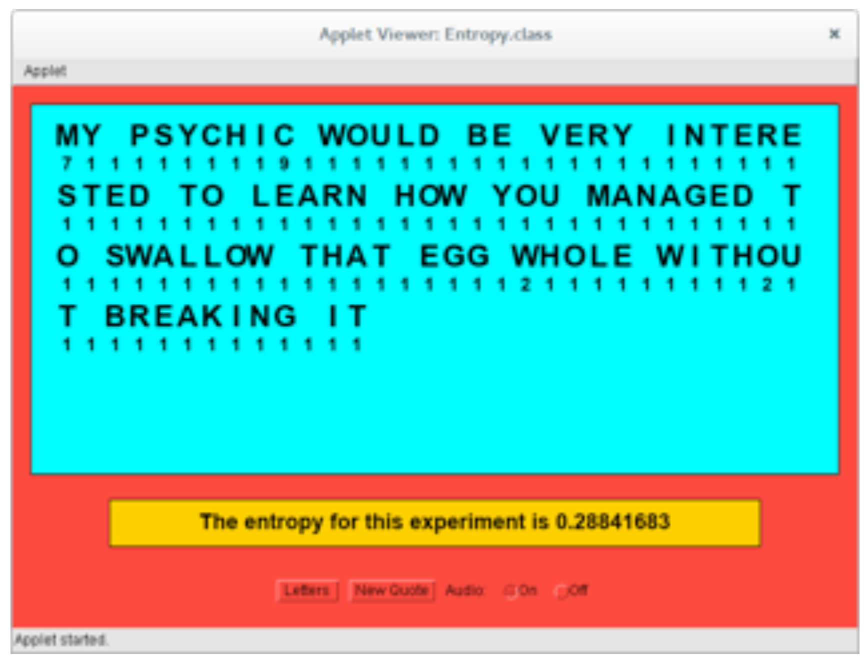


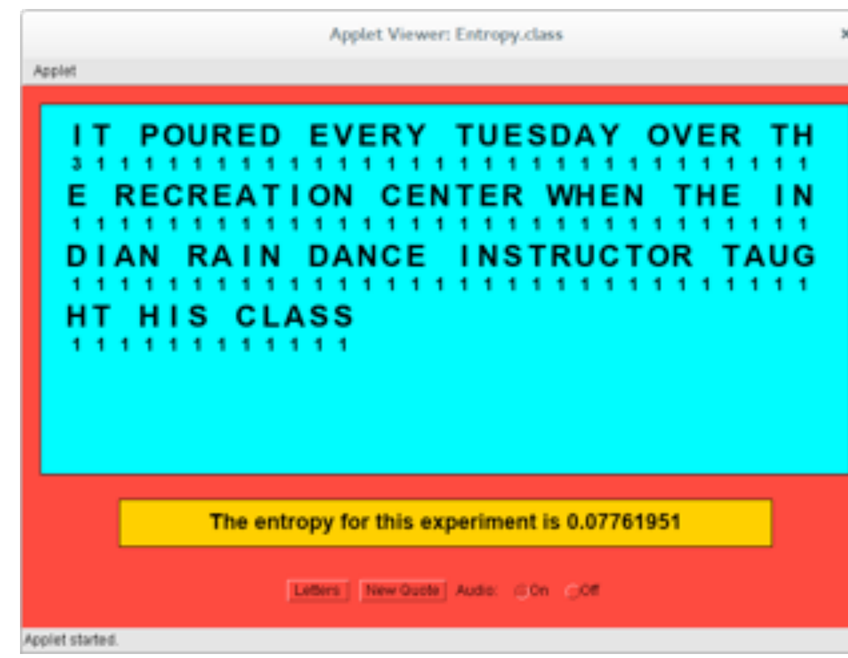
Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

BUT I CAN BEAT YOU ALL!

- guess the next letter; compute entropy (bits per char)
- 0-gram: 4.76,   1-gram: 4.03,   2-gram: 3.32,   3-gram: 3.1
- native speaker: ~1.1 (0.6~1.3);   me: upperbound ~2.3







This Applet only computes Shannon's upperbound!
I'm going to hack it to compute lowerbound as well.

$$\sum_{i=1}^{27} i(q_i^N - q_{i+1}^N)\log i \le F_N \le -\sum_{i=1}^{27} q_i^N \log q_i^N. \qquad (17)$$

Playing Shannon Game: *n*-gram LM

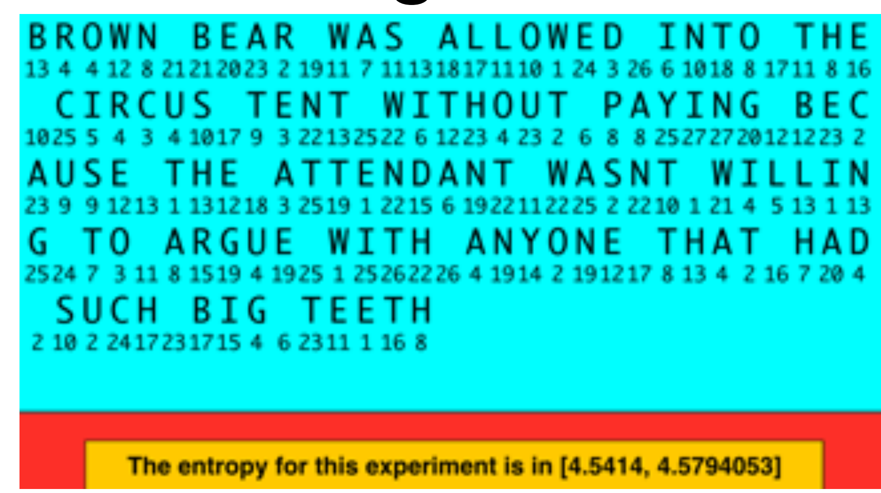
- 0-gram: each char is equally likely (1/27)
- 1-gram: (a) sample from 1-gram distribution from Shakespeare or PTB
- 1-gram: (b) always follow same order: `_ETAIONSRLHDCUMPFGBYWVKXJQZ`
- 2-gram: always follow same order: `Q=>U_A` `J=>UOEAI`

	F_0	F_1	F_2	F_3	F_{word}
26 letter	4.70	4.14	3.56	3.3	2.62
27 letter	4.76	4.03	3.32	3.1	2.14

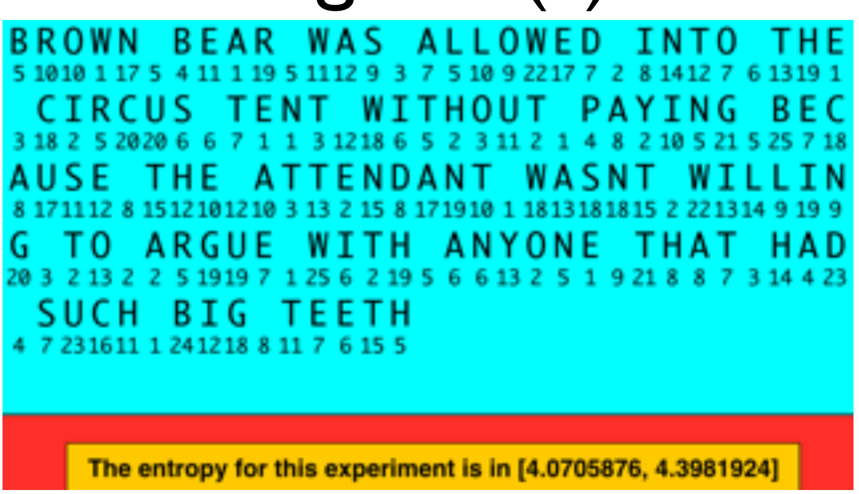
$$\sum_{i=1}^{27} i(q_i^N - q_{i+1}^N)\log i \le F_N \le -\sum_{i=1}^{27} q_i^N \log q_i^N. \qquad (17)$$

Shannon's estimation is less accurate for lower entropy!

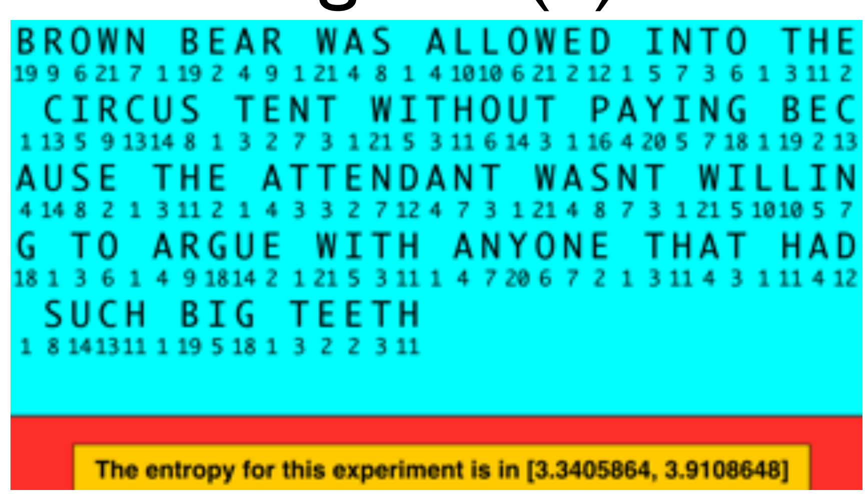
0-gram



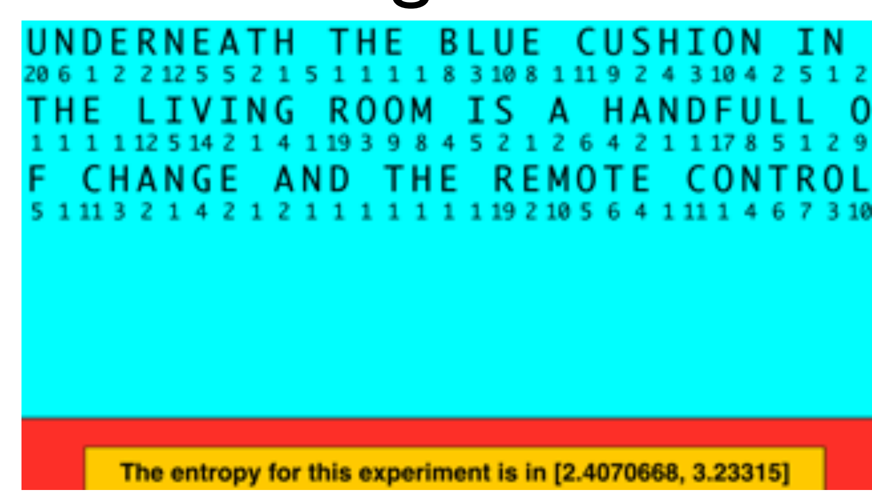
1-gram (a)



1-gram (b)



2-gram



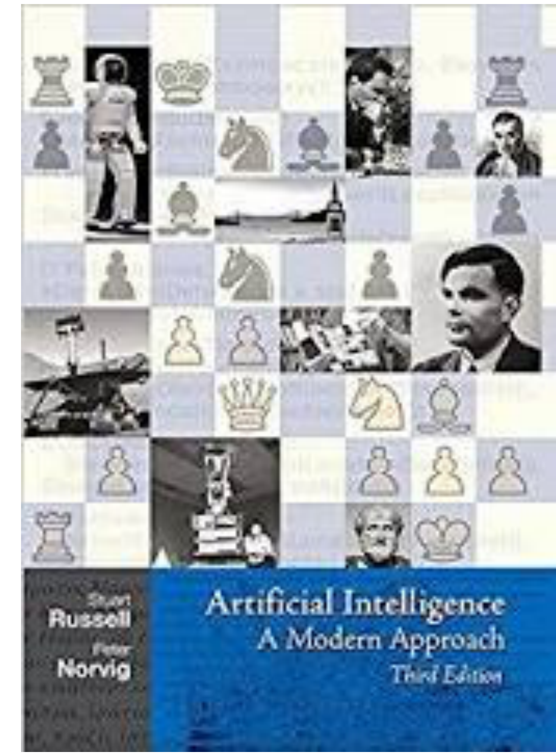
On Google Books (Peter Norvig)

http://norvig.com/mayzner.html

My distillation of the Google books data gives us 97,565 distinct words, which were mentioned 743,842,922,321 times (37 million times more than in Mayzner's 20,000-mention collection). Each distinct word is called a "type" and each mention is called a "token." To no surprise, the most common word is "the". Here are the top 50 words, with their counts (in billions of mentions) and their overall percentage (looking like a Zipf distribution):

WORD	COUNT	PERCENT	bar graph
the	53.10 B	7.14%	the
of	30.97 B	4.16%	of
and	22.63 B	3.04%	and
to	19.35 B	2.60%	to
in	16.89 B	2.27%	in
a	15.31 B	2.06%	a
is	8.38 B	1.13%	is
that	8.00 B	1.08%	that
for	6.55 B	0.88%	for
it	5.74 B	0.77%	it
as	5.70 B	0.77%	as
was	5.50 B	0.74%	was
with	5.18 B	0.70%	with
be	4.82 B	0.65%	be
by	4.70 B	0.63%	by
on	4.59 B	0.62%	on
not	4.52 B	0.61%	not
he	4.11 B	0.55%	he
i	3.88 B	0.52%	i
this	3.83 B	0.51%	this
are	3.70 B	0.50%	are
or	3.67 B	0.49%	or
his	3.61 B	0.49%	his
from	3.47 B	0.47%	from
at	3.41 B	0.46%	at
which	3.14 B	0.42%	which
but	2.79 B	0.38%	but
have	2.78 B	0.37%	have
an	2.73 B	0.37%	an
had	2.62 B	0.35%	had
they	2.46 B	0.33%	they
you	2.34 B	0.31%	you
were	2.27 B	0.31%	were
their	2.15 B	0.29%	their
one	2.15 B	0.29%	one
all	2.06 B	0.28%	all
we	2.06 B	0.28%	we
can	1.67 B	0.22%	can

LET	COUNT	PERCENT	bar graph
E	445.2 B	12.49%	E
T	330.5 B	9.28%	T
A	286.5 B	8.04%	A
O	272.3 B	7.64%	O
I	269.7 B	7.57%	I
N	257.8 B	7.23%	N
S	232.1 B	6.51%	S
R	223.8 B	6.28%	R
H	180.1 B	5.05%	H
L	145.0 B	4.07%	L
D	136.0 B	3.82%	D
C	119.2 B	3.34%	C
U	97.3 B	2.73%	U
M	89.5 B	2.51%	M
F	85.6 B	2.40%	F
P	76.1 B	2.14%	P
G	66.6 B	1.87%	G
W	59.7 B	1.68%	W
Y	59.3 B	1.66%	Y
B	52.9 B	1.48%	B
V	37.5 B	1.05%	V
K	19.3 B	0.54%	K
X	8.4 B	0.23%	X
J	5.7 B	0.16%	J
Q	4.3 B	0.12%	Q
Z	3.2 B	0.09%	Z



1	2grams	3grams	4-grams	5-grams	6-grams	7-grams	8-grams	9-grams
e	th	the	tion	ation	ations	present	differen	different
t	he	and	atio	tions	ration	ational	national	governmen
a	in	ing	that	which	tional	through	consider	overnment
o	er	ion	ther	ction	nation	between	position	formation
i	an	tio	with	other	ection	ication	ifferent	character
n	re	ent	ment	their	cation	differe	governme	velopment
s	on	ati	ions	there	lation	ifferen	vernment	developme
r	at	for	this	ition	though	general	overnmen	evelopmen
h	en	her	here	ement	presen	because	interest	condition
l	nd	ter	from	inter	tation	develop	importan	important
d	ti	hat	ould	ional	should	america	ormation	articular
c	es	tha	ting	ratio	resent	however	formatio	particula
u	or	ere	hich	would	genera	eration	relation	represent
m	te	ate	whic	tiona	dition	nationa	question	individua
f	of	his	ctio	these	ationa	conside	american	ndividual
p	ed	con	ence	state	produc	onsider	characte	relations
g	is	res	have	natio	throug	ference	haracter	political
w	it	ver	othe	thing	hrough	positio	articula	informati
y	al	all	ight	under	etween	osition	possible	nformatio
b	ar	ons	sion	ssion	betwee	ization	children	universit
v	st	nce	ever	ectio	differ	fferent	elopment	following
k	to	men	ical	catio	icatio	without	velopmen	experienc
x	nt	ith	they	latio	people	ernment	developm	stitution
j	ng	ted	inte	about	iffere	vernmen	evelopme	xperience
q	se	ers	ough	count	fferen	overnme	conditio	education
z	ha	pro	ance	ments	struct	governm	ondition	roduction

Bilingual Shannon Game

"From an information theoretic point of view, accurately translated copies of the original text would be expected to contain almost no extra information if the original text is available, so in principle it should be possible to store and transmit these texts with very little extra cost." (Nevill and Bell, 1992)

```
Monolingual Shannon Game (no source sentence)

  I t _ i s _ d e f e n d e d _ t h r o u g h _ r e a s o n i n g .
  D     w     h i m                             e i f i     a     ,
  m     t                                       a   s       _     _
  c     o                                       n
                                                c
                                                i
                                                f
                                                m
                                                o
                                                d
                                                p
                                                t

Bilingual Shannon Game (source sentence = "Se defiende con argumentos.")

  I t _ i s _ d e f e n d e d _ t h r o u g h _ r e a s o n i n g .
                                w               d           .
                                                a
```

If I am fluent in Spanish, then English translation adds no new info.

If I understand 50% Spanish, then English translation adds some info.

If I don't know Spanish at all, then English should be have the same entropy as in the monolingual case.

Entropy **2017**, *19*(1), 15; doi:10.3390/e19010015

Humans Outperform Machines at the Bilingual Shannon Game

Marjan Ghazvininejad[†,*] and Kevin Knight[†]

http://www.mdpi.com/1099-4300/19/1/15

Other Resources

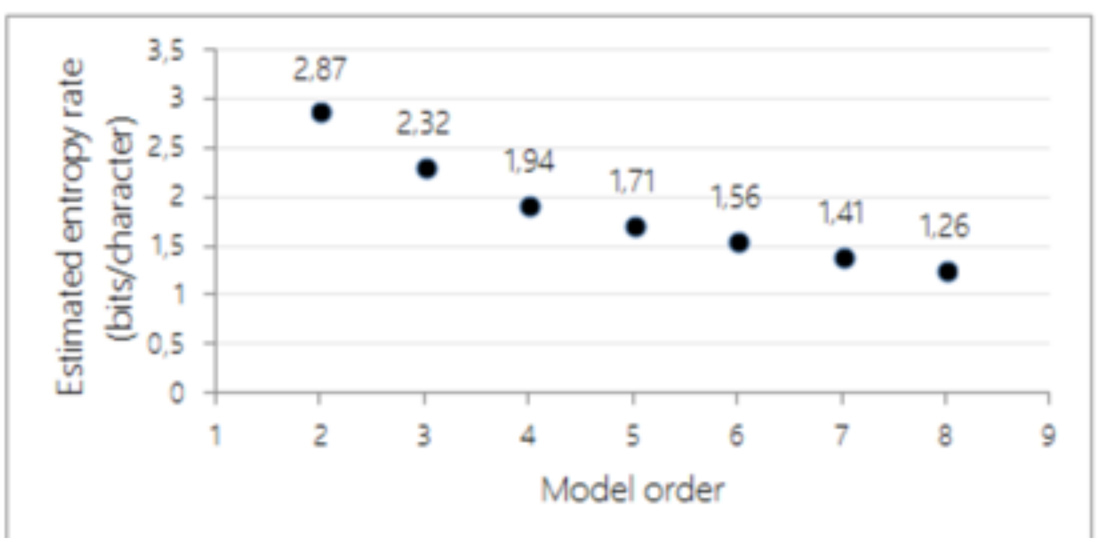
- "Unreasonable Effectiveness of RNN" by Karpathy
- Yoav Goldberg's follow-up for n-gram models (ipynb)

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

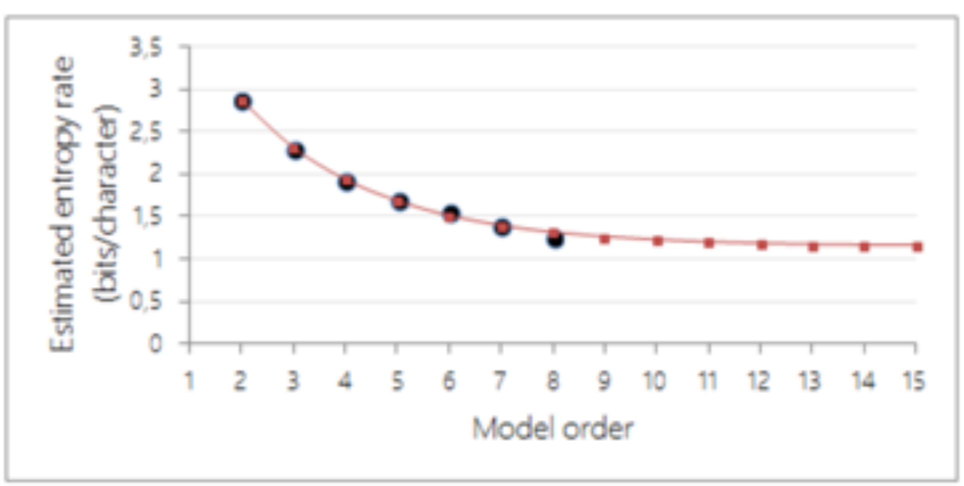
http://nbviewer.jupyter.org/gist/yoavg/d76121dfde2618422139

http://pit-claudel.fr/clement/blog/an-experimental-estimation-of-the-entropy-of-english-in-50-lines-of-python-code/

Running this algorithm on the entire Open American National Corpus (about 95 million characters) yields the following results:



As a rough example, call this sequence of values F_k and assume that it verifies the recurrence equation $F_{k+1} - F_k = \alpha(F_n - F_{n-1})$. Then the α that yields the best approximation (taking the two initial values for granted since they are less likely to suffer from sampling errors) is $\alpha \approx 0.68$ ($\mathcal{L}^2$ error: $6.7 \cdot 10^{-3}$), and the corresponding entropy rate is $h \approx 1.14$ bits/character.



Extrapolated entropy rate values for $\alpha \approx 0.68$. In this heuristic model, the limit entropy rate is $h \approx 1.14$ bits/character.