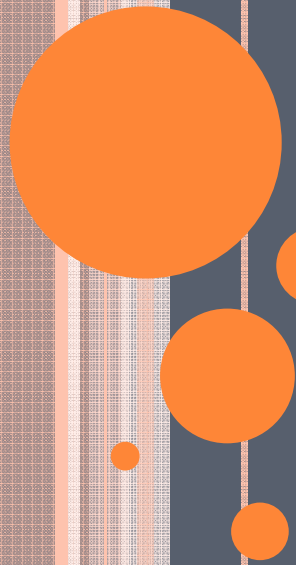
ENTROPY IN NLP

Presented by
Avishek Dan (113050011)
Lahari Poddar (113050029)
Aishwarya Ganesan (113050042)

Guided by
Dr. Pushpak Bhattacharyya



Motivation


RED
SEEDLESS
RAISES
JACKPOT
$8,850
THINGS
$4,000
TURNS: 1
JOEL: $2,800
The Wheel landed on $550. Choose a consonant.
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
SPIN
BUY VOWEL
SOLVE
PAUSE

- Aoccdrnig to rseearch at an Elingsh uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoatnt tihng is that the frist and lsat ltteer is at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by it slef but the wrod as a wlohe.

40% Removed

Dark s r c res ow e n eit er side the froze w erwa . T
trees h d b n tr p d b c t w n o heir white c v ring of
f s and h y s med e n towa ds e c o her la k and
i o s, i e f l gh . A as ilence reig ed o he
and. The lan e f a a de l tion, life e s, i
mo ment, so n nd cold h the pi t o it s n t ev n hat
o sa nes . h e wa hin n it of laughter, bu a lau h r
e t rib e a dn - laught r a was mi hle s as
the i e of t e s a la hte ol as he ro t a d ar ak n
f th g mne of inf llib i . It wa he mas erful n
incom un b e w s om of t rnity l ugh n at e lity f fe
nd e effort f ife. I as t e , savage, fro n-
h a r hland Wi d.

30% Removed

Da s ruce fores r on ith r i e the froz n ater ay The
t s ha een st ippe b re e wi d f heir white c vering of
ost, an they e d o ean towards eac othe , lac a d
om nou , n t e f in ight. A vast il nc re ned o e he
la d. The l d s l wa d s a ion, if s, without
m men o e and old that t e pi t of it was not e en tha
o sadnes T ere s a hint in t of laug r, ut o a laugh e
o er ble t n a y sadn s a la ghte at as irthless s
the mil o h phi x, a ghter cold as the fro t a d par ak ng
f the grim es of n al b l ty was th m s r ul and
co mu i able wi m of ter la ghi t t e futil ty o if
an the ef rt of li e. It as the Wild, h avage, froze -
hearte Northland Wil .

20% Removed

Dark spruce forest frowned n either side the roze ate wa . The trees had ee stripp d by ecent wi d f thei white coverin of frost d they eemed to lean towa ds e ch o h r, bl ck and mino s, in the ad ng l ght. A vas sil n e reigned over the land. The land i sel was e olatio , ifeless, wit out movem n , s lon n cold ha e spi it of s n t eve hat adn s. The e was hi i it of a ght , but f a laughter ore ible t an any s ne s - a laughter t a was mi hless as he mile of he sphinx, a aug ter col as the f ost nd arta in of th grimn ss of i fall bility It as the asterful and inc mmunicabl wisdo of e ernity ugh ng at the futilit of li and t e for of ife. It w the Wild, the avage, rozen-hea t d N rthla d W ld.

10% Removed

Dark s ru e forest frowned on either side the frozen waterw y. The trees had bee stripped by a recent ind of t eir w ite covering of rost, and they seemed o lean towards each ot er, black and ominous, in the fading li h . A vast silence reigned ver the land The land itself w s a deso ation, lifel ss, without ovement, s l n and cold hat the spiri of it wa not even that of sa ness. here was a hint in it of laughte , but of a l ug ter more terrible than any sadness - a ughte that as mirthless s he smile of the sphinx a laug ter cold as the rost a d p rt k ng of the grimness of infal i ility. It as the masterful and incommunica le wisdom of eternity laughing at th futility of life an the effort f life. It was e Wild, the sa ag , froz n- earte Northland Wild.

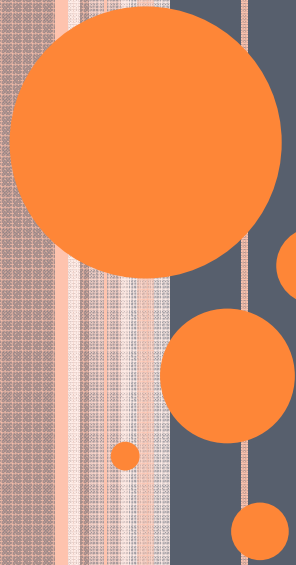
0% Removed

Dark spruce forest frowned on either side the frozen waterway. The trees had been stripped by a recent wind of their white covering of frost, and they seemed to lean towards each other, black and ominous, in the fading light. A vast silence reigned over the land. The land itself was a desolation, lifeless, without movement, so lone and cold that the spirit of it was not even that of sadness. There was a hint in it of laughter, but of a laughter more terrible than any sadness - a laughter that was mirthless as the smile of the sphinx, a laughter cold as the frost and partaking of the grimness of infallibility. It was the masterful and incommunicable wisdom of eternity laughing at the futility of life and the effort of life. It was the Wild, the savage, frozen-hearted Northland Wild.

From Jack London's "White Fang"

IN A previous paper[1] the entropy and redundancy of a language ha
been defined. The entropy is a statistical parameter which measure
in a certain sense, how much information is produced on the average f
each letter of a text in the language. If the language is translated into bina
digits (0 or 1) in the most efficient way, the entropy H is the average numb
of binary digits required per letter of the original language. The redundanc
on the other hand, measures the amount of constraint imposed on a text
the language due to its statistical structure, e.g., in English the high fr
quency of the letter E, the strong tendency of H to follow T or of U to follo
Q. It was estimated that when statistical effects extending over not mo
than eight letters are considered the entropy is roughly 2.3 bits per lette
the redundancy about 50 per cent.

Since then a new method has been found for estimating these quantitie
which is more sensitive and takes account of long range statistics, influenc
extending over phrases, sentences, etc. This method is based on a study
the predictability of English; how well can the next letter of a text be pr
dicted when the preceding N letters are known. The results of some exper
ments in prediction will be given, and a theoretical analysis of some of th
properties of ideal prediction. By combining the experimental and theoret
cal results it is possible to estimate upper and lower bounds for the entrop
and redundancy. From this analysis it appears that in ordinary literar



Entropy

ENTROPY

- Entropy or self-information is the average uncertainty of a single random variable:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

  (i) H(x) >=0,

  (ii) H(X) = 0 only when the value of X is determinate, hence providing no new information

- From a language perspective, it is the information that is produced on the average for each letter of text in the language

EXAMPLE: SIMPLE POLYNESIAN

- Random sequence of letters with probabilities:

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

- Per-letter entropy is:

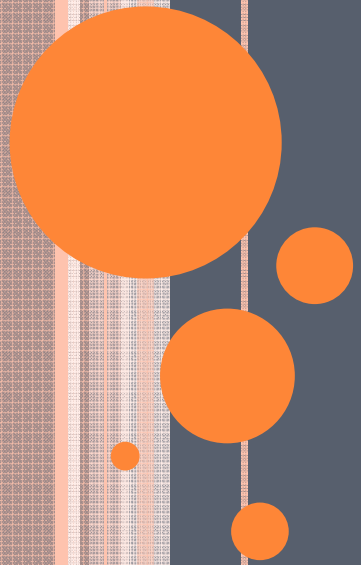
$$H(P) = -\sum_{i \in \{p,t,k,a,i,u\}} P(i) \log P(i)$$

$$= -\left[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}\right]$$

$$= 2\frac{1}{2} \text{ bits}$$

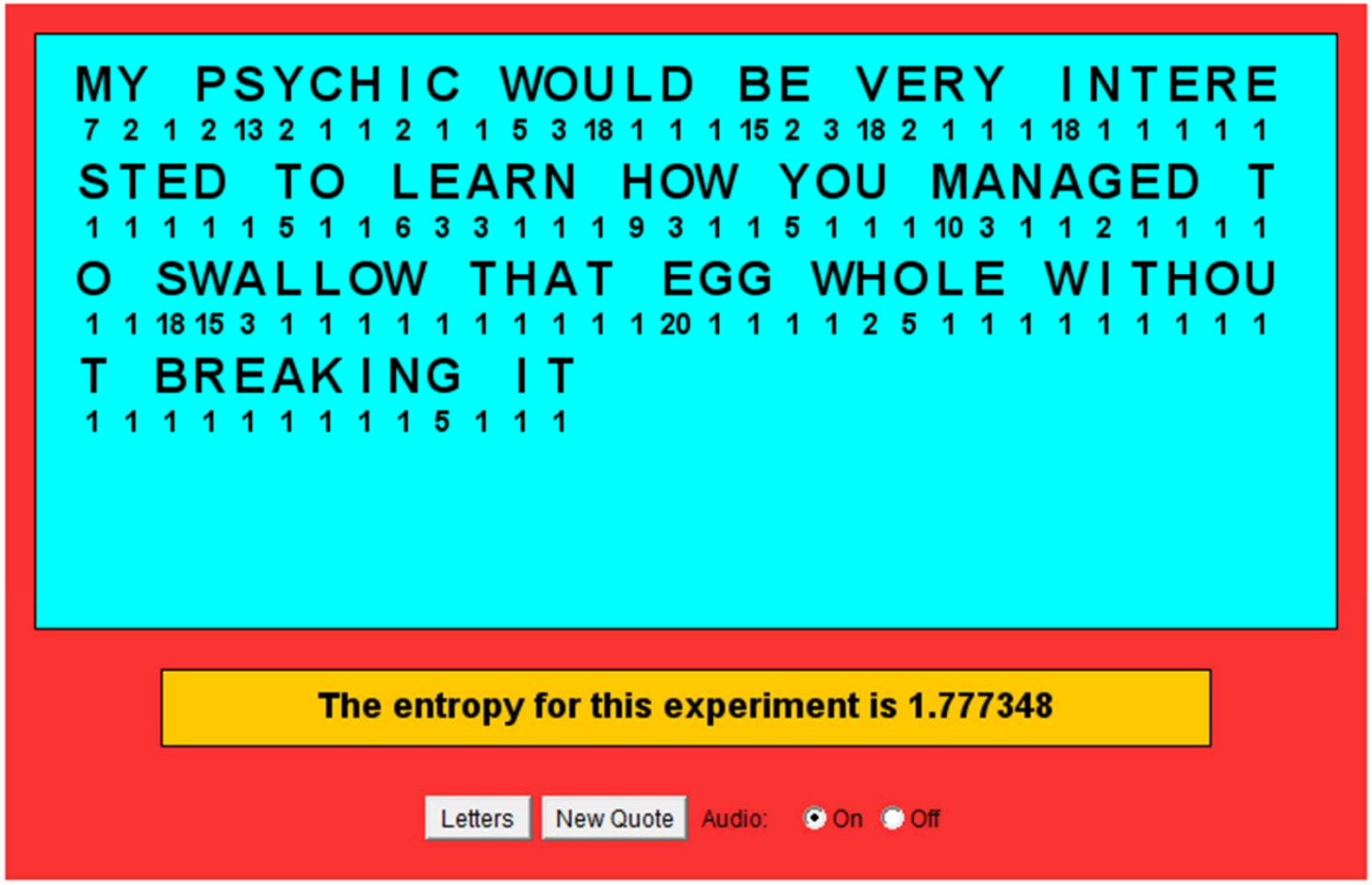
- Code to represent language:

p	t	k	a	i	u
100	00	101	01	110	111



Shanon's Experiment to determine entropy

Shannon's Entropy Experiment



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

Calculation of Entropy

- User as a language model
- Encode number of guesses required
- Apply entropy encoding algorithm (lossless compression)

ENTROPY OF A LANGUAGE

- Series of approximations

  $F_0, F_1, F_2 \ldots F_n$

$$F_N = -\sum_{i,j} p(b_i, j) \log_2 p_{b_i}(j)$$

$$= -\sum_{i,j} p(b_i, j) \log_2 p(b_i, j) + \sum_i p(b_i) \log_2 p(b_i)$$

$$H = \lim_{n \to \infty} F_N$$

ENTROPY OF ENGLISH

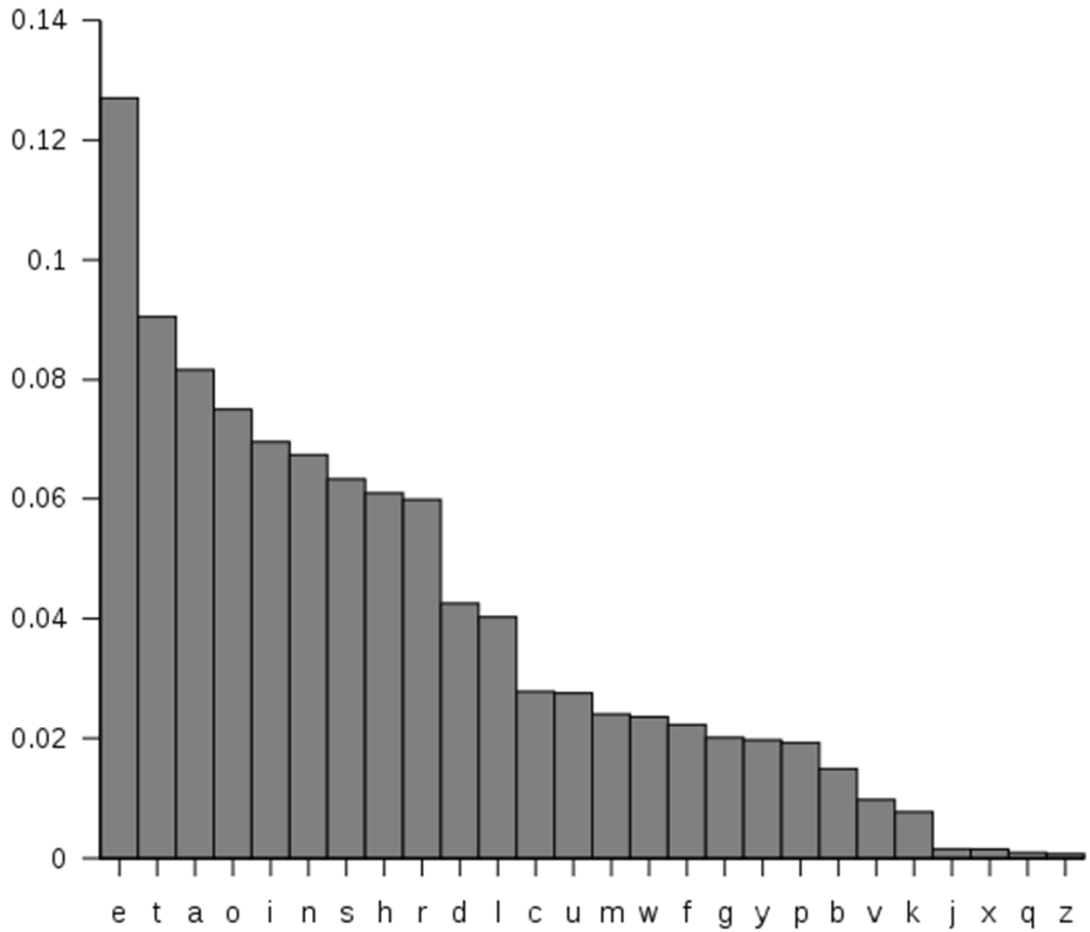
- $F_0 = \log_2 26 = 4.7$ bits per letter

- $$F_1 = -\sum_{i=1}^{26} p(i) \log_2 p(i) = 4.14 \text{ bits per letter}$$

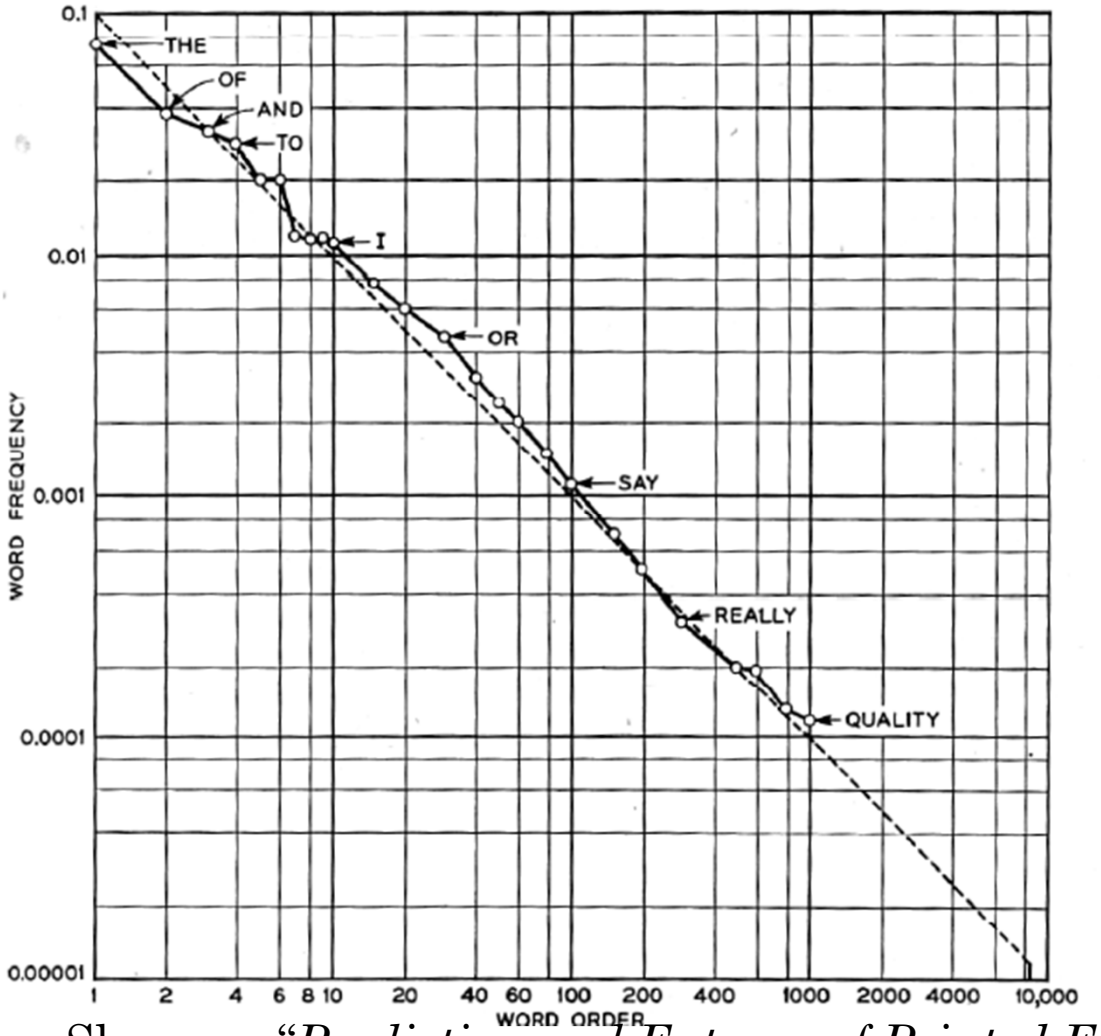
- $$F_2 = -\sum_{i,j} p(i,j) \log_2 p_i(j)$$
$$= -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p(i) \log_2 p(i)$$
$$= 7.70 - 4.14 = 3.56 \text{ bits per letter.}$$

- $$F_3 = -\sum_{i,j,k} p(i,j,k) \log_2 p_{ij}(k)$$
$$= -\sum_{i,j,k} p(i,j,k) \log_2 p(i,j,k) + \sum_{i,j} p(i,j) \log_2 p(i,j)$$
$$\doteq 11.0 - 7.7 = 3.3$$

Relative frequency of occurrence of English alphabets



Word Entropy of English

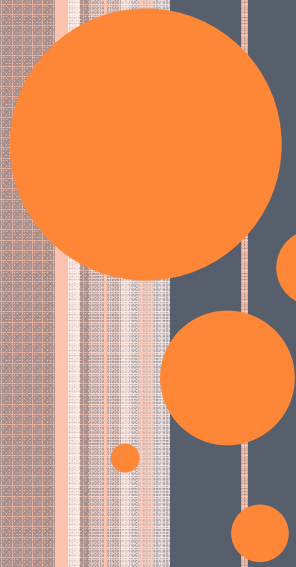


Source: Shannon "*Prediction and Entropy of Printed English*"

Zipf's Law

- Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table

$$p_n = \frac{.1}{n}$$



Language modeling

Language Modeling

A language model computes either:

- probability of a sequence of words:

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$$

- probability of an upcoming word:

$$P(w_n | w_1, w_2 \ldots w_{n-1})$$

Applications of Language Models

- POS Tagging
- Machine Translation
  - P(*heavy* rains tonight) > P(*weighty* rains tonight)
- Spell Correction
  - P(about fifteen *minutes* from) > P(about fifteen *minuets* from)
- Speech Recognition
  - P(I saw a van) >> P(eyes awe of an)

N-GRAM MODEL

Chain rule

$$P(w_1 w_2 \rightleftharpoons w_n) = \prod_i P(w_i \mid w_1 w_2 \rightleftharpoons w_{i-1})$$

Markov Assumption

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1})$$

Maximum Likelihood Estimate (for k=1)

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

EVALUATION OF LANGUAGE MODELS

A good language model gives a high probability to real English

- Extrinsic Evaluation
  - For comparing models A and B
  - Run applications like POSTagging, translation in each model and get an accuracy for A and for B
  - Compare accuracy for A and B

- Intrinsic Evaluation
  - Use of cross entropy and perplexity
  - True model for data has the lowest possible entropy / perlexity

RELATIVE ENTROPY

- For two probability mass functions, p(x), q(x) their relative entropy:

$$D(\mathrm{p} \parallel \mathrm{q}) = \sum_{x \in \mathcal{X}} \mathrm{p}(x) \log \frac{\mathrm{p}(x)}{\mathrm{q}(x)}$$

- Also known as *KL divergence*, a measure of how different two distributions are.

$$D(\mathrm{p} \parallel \mathrm{q}) = E_{\mathrm{p}} \left(\log \frac{\mathrm{p}(X)}{\mathrm{q}(X)} \right)$$

CROSS ENTROPY

- Entropy as a measure of how surprised we are, measured by pointwise entropy for model m:

  H(w/h) = - log(m(w/h))

- Produce q of real distribution to minimize D(p||q)

- The cross entropy between a random variable X with p(x) and another q (a model of p) is:

$$H(X,q) = H(X) + D(p \parallel q)$$
$$= -\sum_x p(x) \log q(x)$$
$$H(L,m) \approx -\frac{1}{n} \log m(x_{1n})$$

PERPLEXITY

- Perplexity is defined as

$$PP(x_{1n}, m) = 2^{H(x_{1n}, m)}$$

- Probability of the test set assigned by the language model, normalized by the number of word

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

- Most common measure to evaluate language models

EXAMPLE

Training set

there is a big house
i buy a house
they buy the new house

Model

$p(big\|a) = 0.5$	$p(is\|there) = 1$	$p(buy\|they) = 1$
$p(house\|a) = 0.5$	$p(buy\|i) = 1$	$p(a\|buy) = 0.5$
$p(new\|the) = 1$	$p(house\|big) = 1$	$p(the\|buy) = 0.5$
$p(a\|is) = 1$	$p(house\|new) = 1$	$p(they\|< s >) = .333$

Test sentence S: *they buy a big house*

$$p(S) = \underbrace{0.333}_{they} \times \underbrace{1}_{buy} \times \underbrace{0.5}_{a} \times \underbrace{0.5}_{big} \times \underbrace{1}_{house} = 0.0833$$

EXAMPLE (CONTD.)

- Cross Entropy:

$$
\begin{aligned}
H(p,m) &= -\frac{1}{5}\log p(S) \\
&= -\frac{1}{5}(\underbrace{\log 0.333}_{they} + \underbrace{\log 1}_{buy} + \underbrace{\log 0.5}_{a} + \underbrace{\log 0.5}_{big} + \underbrace{\log 1}_{house}) \\
&= -\frac{1}{5}(\underbrace{-1.586}_{they} + \underbrace{0}_{buy} + \underbrace{-1}_{a} + \underbrace{-1}_{big} + \underbrace{0}_{house}) = 0.7173
\end{aligned}
$$

- Perplexity:

$$PP = 1.6441$$

SMOOTHING

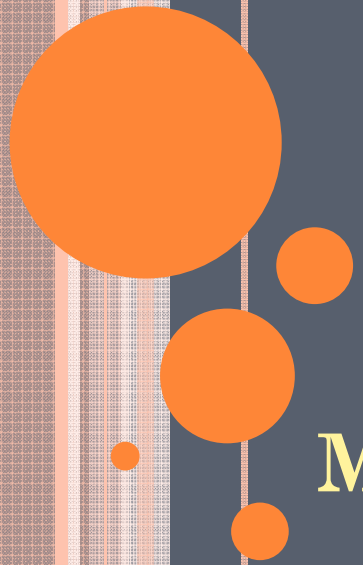
- Bigrams with zero probability - cannot compute perplexity
- When we have sparse statistics Steal probability mass to generalize better ones.
- Many techniques available
- Add-one estimation : add one to all the counts

$$P_{Add-1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)+1}{c(w_{i-1})+V}$$

PERPEXITY FOR N-GRAM MODELS

- Perplexity values yielded by n-gram models on English text range from about 50 to almost 1000 (corresponding to cross entropies from about 6 to 10 bits/word)
- Training: 38 million words from WSJ by Jurafsky
- Vocabulary: 19,979 words
- Test: 1.5 million words from WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109



Maximum Entropy Model

Statistical Modeling

- Constructs a stochastic model to predict the behavior of a random process
- Given a sample, represent a process
- Use the representation to make predictions about the future behavior of the process

- Eg: Team selectors employ batting averages, compiled from history of players, to gauge the likelihood that a player will succeed in the next match. Thus informed, they manipulate their lineups accordingly

Stages Of Statistical Modeling

1. Feature Selection :Determine a set of statistics that captures the behavior of a random process.

2. Model Selection: Design an accurate model of the process--a model capable of predicting the future output of the process.

Motivating Example

- Model an expert translator's decisions concerning the proper French rendering of the English word *on*
- A model(p) of the expert's decisions assigns to each French word or phrase(*f*) an estimate, *p*(*f*), of the probability that the expert would choose *f* as a translation of *on*
- Our goal is to
  - Extract a set of facts about the decision-making process from the sample
  - Construct a model of this process

Motivating Example

- A clue from the sample is the list of allowed translations
  - *on* → {sur, dans, par, au bord de}
- With this information in hand, we can impose our first constraint on p:

$$p(sur) + p(dans) + p(par) + p(au\ bord\ de) = 1$$

- The most uniform model will divide the probability values equally
- Suppose we notice that the expert chose either *dans* or *sur* 30% of the time, then a second constraint can be added

$$p(dans) + p(sur) = 3/10$$

- Intuitive Principle: Model all that is known and assume nothing about that which is unknown

Maximum Entropy Model

- A random process which produces an output value y, a member of a finite set $\boldsymbol{Y}$.
  - $y \in$ {sur, dans, par, au bord de}
- The process may be influenced by some contextual information x, a member of a finite set $\boldsymbol{X}$.
  - x could include the words in the English sentence surrounding *on*
- A stochastic model: accurately representing the behavior of the random process
  - Given a context x, the process will output y

Maximum Entropy Model

- Empirical Probability Distribution:

$$\tilde{p}(x, y) \equiv \frac{1}{N} \times \text{number of times that } (x, y) \text{ occurs in the sample}$$

- Feature Function:

$$f(x, y) = \begin{cases} 1 & \text{if } y = sur \text{ and } hold \text{ follows } on \\ 0 & \text{otherwise} \end{cases}$$

- Expected Value of the Feature Function
  - For training data:

$$\tilde{p}(f) \equiv \sum_{x, y} \tilde{p}(x, y) f(x, y) \qquad (1)$$

  - For model:

$$p(f) \equiv \sum_{x, y} \tilde{p}(x) p(y \mid x) f(x, y) \qquad (2)$$

Maximum Entropy Model

- To accord the model with the statistic, the expected values are constrained to be equal

$$p(f) = \tilde{p}(f) \qquad (3)$$

- Given n feature functions f_i, the model p should lie in the subset $\boldsymbol{C}$ of $\boldsymbol{P}$ defined by

- Choose the model p* with maximum entropy H(p):

$$\boldsymbol{C} \equiv \{p \in \boldsymbol{P} \mid p(f_i) = \tilde{p}(f_i) \text{ for } i \in \{1,2,\ldots,n\}\} \qquad (4)$$

$$H(p) = -\sum_{x,y} \tilde{p}(x)p(y \mid x)\log p(y \mid x) \qquad (5)$$

APPLICATIONS OF MEM

Statistical Machine Translation

- A French sentence F, is translated to an English sentence E as:

$$E = \arg \max_{E} p(F \mid E) P(E)$$

- Addition of MEM can introduce context-dependency:
  - $P_e(f \mid x)$: Probability of choosing e (English) as the rendering of f (French) given the context x

Translation	e_{-3}	e_{-2}	e_{-1}		e_{+1}	e_{+2}	e_{+3}
dans	*the*	*committee*	*stated*	·	*a*	*letter*	*to*
à	*work*	*was*	*required*	·	*respect*	*of*	*the*
au cours de				·	*the*	*fiscal*	*year*
dans	*by*	*the*	*government*	·	*the*	*same*	*postal*
à	*of*	*diphtheria*	*reported*	·	*Canada*	*,*	*by*
de	*not*	*given*	*notice*	·	*the*	*ordinary*	*way*

PART OF SPEECH TAGGING

- The probability model is defined over HxT

  H : set of possible word and tag contexts(histories)

  T : set of allowable tags

- Entropy of the distribution :

$$H(p) = -\sum_{h \in H, t \in T} p(h,t) \log p(h,t)$$

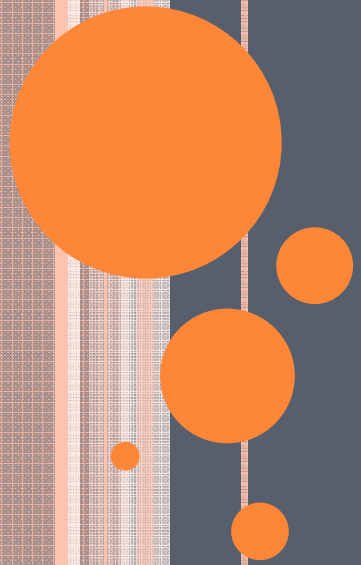
- Sample Feature Set :

Condition	Features	
w_i is not rare	$w_i = X$	& $t_i = T$
w_i is rare	X is prefix of w_i, $\|X\| \leq 4$	& $t_i = T$
	X is suffix of w_i, $\|X\| \leq 4$	& $t_i = T$
	w_i contains number	& $t_i = T$
	w_i contains uppercase character	& $t_i = T$
	w_i contains hyphen	& $t_i = T$
$\forall\ w_i$	$t_{i-1} = X$	& $t_i = T$
	$t_{i-2}t_{i-1} = XY$	& $t_i = T$
	$w_{i-1} = X$	& $t_i = T$
	$w_{i-2} = X$	& $t_i = T$
	$w_{i+1} = X$	& $t_i = T$
	$w_{i+2} = X$	& $t_i = T$

- Precision :96.6%

Preposition Phrase Attachment

- MEM produces a probability distribution for the PP-attachment decision using only information from the verb phrase in which the attachment occurs
- conditional probability of an attachment is $p(d \mid h)$
  - *h is the history*
  - d∈{0, 1} , corresponds to a noun or verb attachment (respectively)
- Features: Testing for features should only involve
  - Head Verb (V)
  - Head Noun (N1)
  - Head Preposition (P)
  - Head Noun of the Object of the Preposition (N2)
- Performance :
  - Decision Tree : 79.5 %
  - MEM : 82.2%



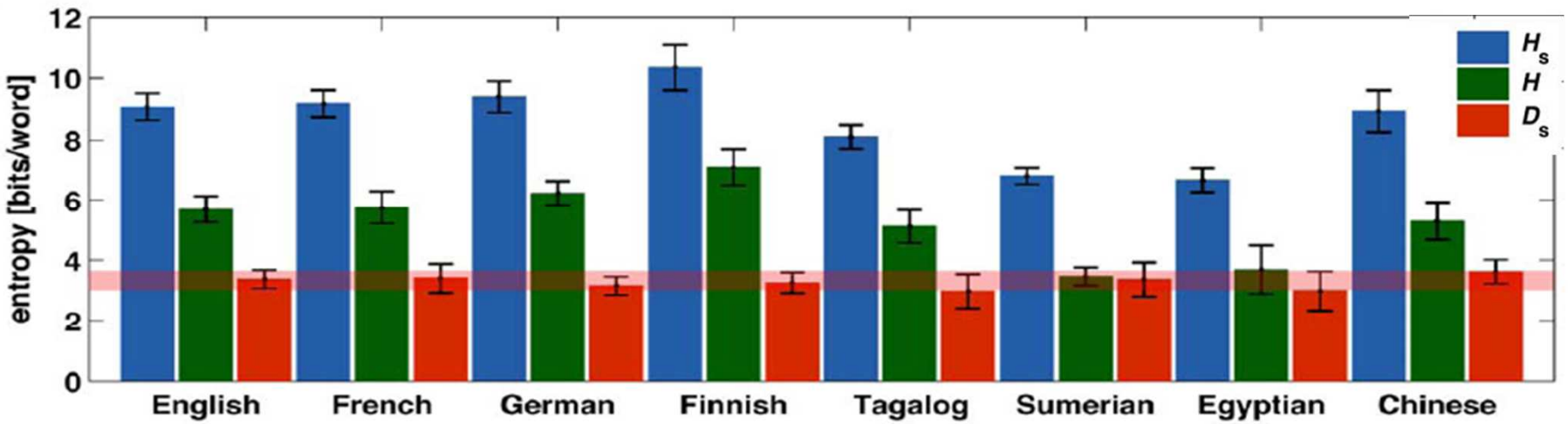
Entropy of other languages

ENTROPY OF EIGHT LANGUAGES BELONGING TO FIVE LINGUISTIC FAMILIES

- Indo-European: English, French, and German
- Finno-Ugric: Finnish
- Austronesian: Tagalog
- Isolate: Sumerian
- Afroasiatic: Old Egyptian
- Sino-Tibetan: Chinese

Hs- entropy when words are random , H- entropy when words are ordered

Ds= Hs-H



Source: Universal Entropy of Word Ordering Across Linguistic Families

Entropy of Hindi

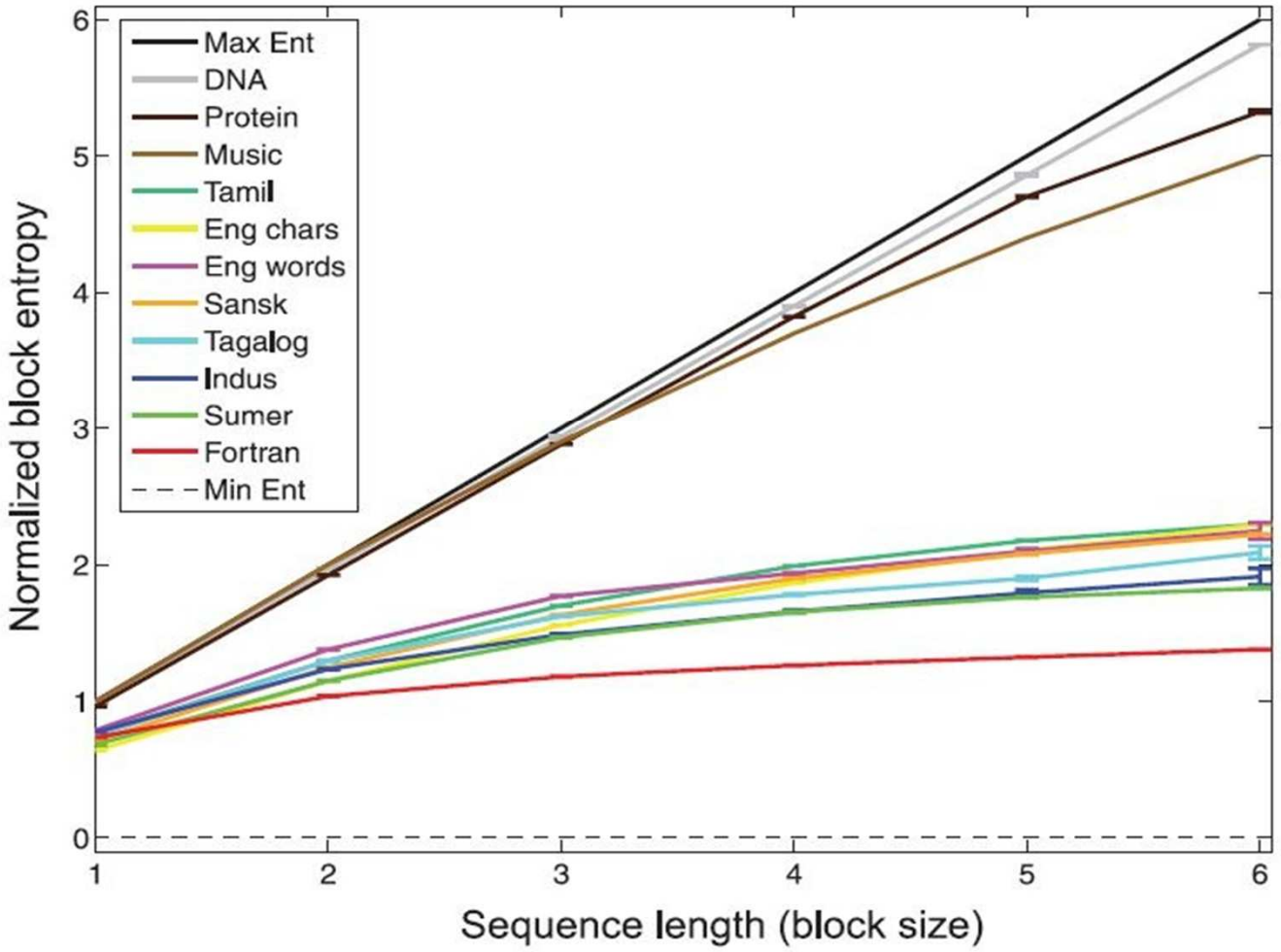
- *Zero-order :* 5.61 bits/symbol.
- *First-order :* 4.901 bits/symbol.
- *Second-order :* 3.79 bits/symbol.
- *Third-order :* 2.89 bits/symbol.
- *Fourth-order :* 2.31 bits/symbol.
- *Fifth-order :* 1.89 bits/symbol.

ENTROPY TO PROVE THAT A SCRIPT REPRESENT LANGUAGE

- Pictish (a Scottish, Iron Age culture) symbols revealed as a written language through application of Shannon entropy

- In Entropy, the Indus Script, and Language by proving the block entropies of the Indus texts remain close to those of a variety of natural languages and far from the entropies for unordered and rigidly ordered sequences

Entropy of linguistic and non-linguistic languages



Source: *Entropy, the Indus Script, and Language*

Conclusion

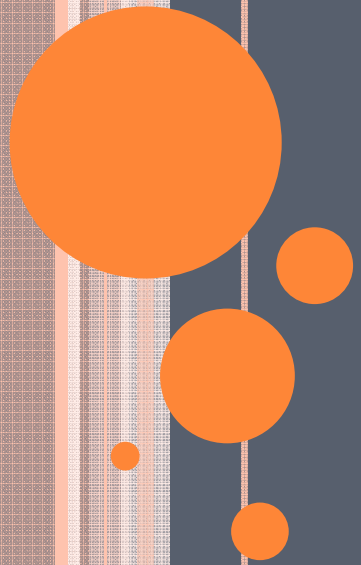
- The concept of entropy dates back to biblical times but it has wide range of applications in modern NLP

- NLP is heavily supported by Entropy models in various stages. We have tried to touch upon few aspects of it.

References

- C. E. Shannon, *Prediction and Entropy of Printed English,* Bell System Technical Journal Search, Volume 30, Issue 1, January 1951
- http://www.math.ucsd.edu/~crypto/java/ENTROPY/
- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.
- Daniel Jurafsky and James H. Martin, SPEECH and LANGUAGE PROCESSING An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition ,January 6, 2009
- Marcelo A. Montemurro, Damia´n H. Zanette, *Universal Entropy of Word Ordering Across Linguistic Families,* PLoS ONE 6(5): e19875. doi:10.1371/journal.pone.0019875

- Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, Ronojoy Adhikari, Iravatham Mahadevan, *Entropy, the Indus Script, and Language: A Reply to R. Sproat*
- Adwait Ratnaparkhi, *A Maximum Entropy Model for Prepositional Phrase Attachment*, HLT '94
- Berger, A Maximum Entropy Approach to Natural Language Processing, 1996
- Adwait Ratnaparkhi, A Maximum Entropy Model for Part-Of-Speech Tagging, 1996



THANK YOU