Information Theoretic Approaches to the Study of Language

Matthew W. Crocker
Vera Demberg

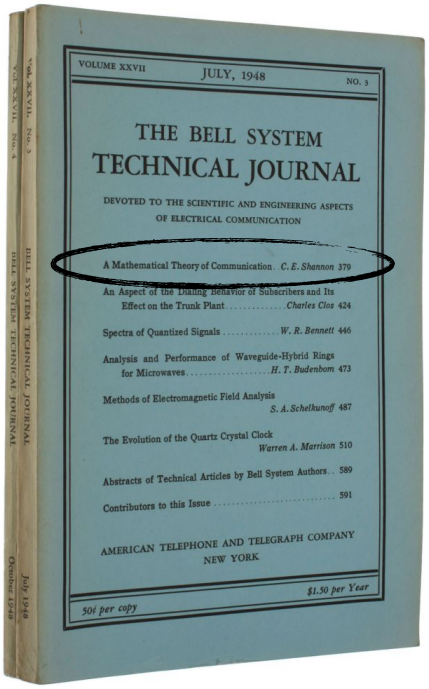
Summer 2017

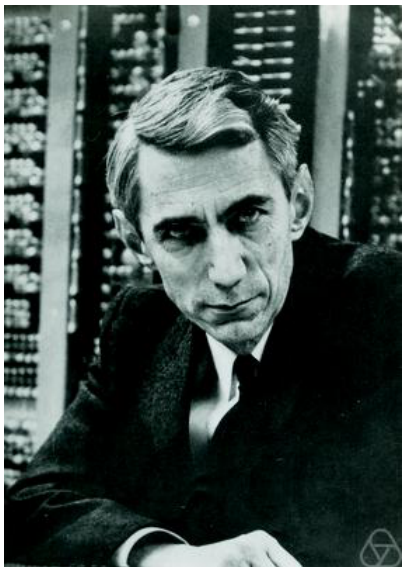
Communication

Language as Information

- We can think of language as a communication system, in which information is transmitted from speaker to hearer
- *Rationality* suggests that language, and language use, will be optimized to transmit information efficiently and accurately
- Information Theory is a rational mathematical theory for the efficient transmission of information across an imperfect “noisy” channel.

1948





Claude Shannon

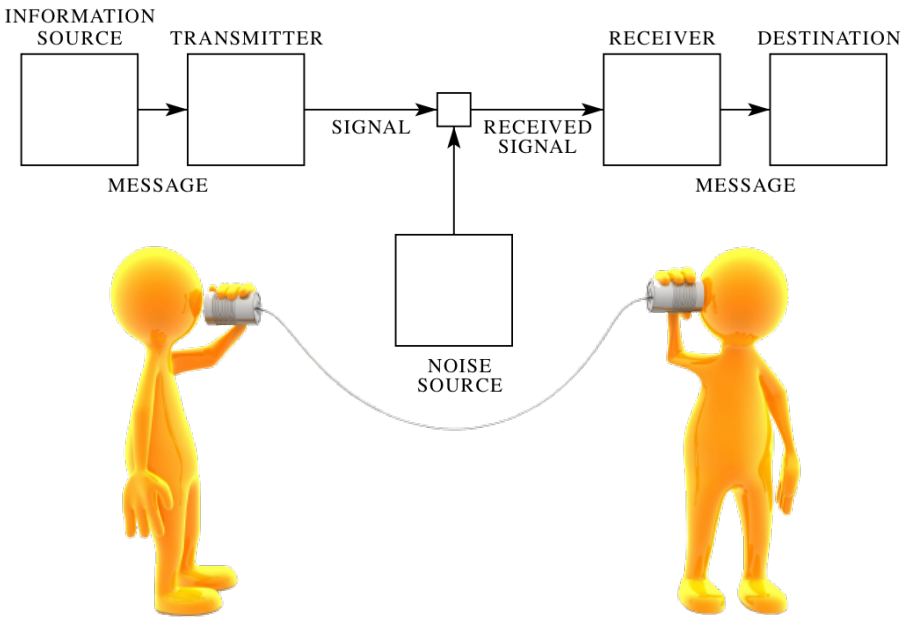
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.



This Course

- What is Information Theory?
- What does it tell us about rational communication?
- Can and should these ideas inform theories of human communication and language?
- What do information theoretic models of language look like?
- How can we empirically investigate these ideas?

Assessment

- Students will form groups to prepare a research proposal of what linguistic phenomenon could be investigated using surprisal or the UID hypothesis:
  - propose what kind of research method to use for tackling the question
  - what results they would expect,
  - and what it would mean to find different results than the expected ones.
- Proposals will presented as posters, which will be present to the rest of the course at a time slot roughly two weeks after the end of the course
- Form groups by tomorrow, and come up with a rough topic or two by Thurs. We'll discuss the proposal during the tutorial slot.

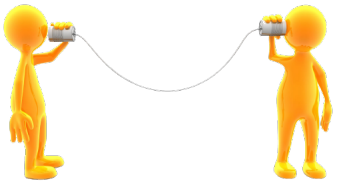
What is Information

How much information is conveyed by a particular message, event, outcome?

The minimal unit of information we can convey is the outcome of a binomial event: yes/no, 1/0 …

- So it's useful to consider this unit, *bits*, as the the basic unit of measure

Information



- A measure of the disorder or predictability in a system:

- What if there are two coins?

  - The (average) number of yes/no questions needed to completely specify the state of a system

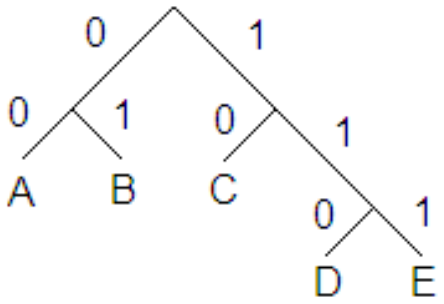
Number of Questions in General?

Number of States	Number of Questions
2 states	1 question
4 states	2 questions
8 states	3 questions
16 states	4 questions

$\log_2$(# of states) = number of yes-no questions

Binary Coding Trees

- Suppose we have 5 messages
  - $\log_2 5 = 2.32$ bits
  - but we can't devise a code that achieves this



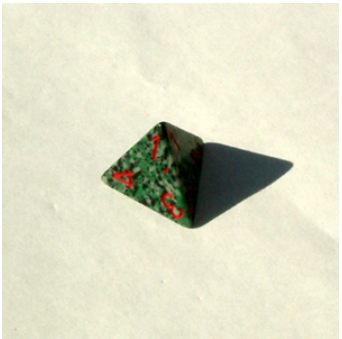
What's the average # of bits required to send a letter?

$$Avg = \frac{(3*2)+(2*3)}{5} = \frac{12}{5} = 2.4 bits$$

Consider Dice

For each Die

$H = \log_2(6)$
$= 2.585$ bits



$H = \log_2(4)$
$= 2.000$ bits

$H = \log_2(20)$
$= 4.322$ bits

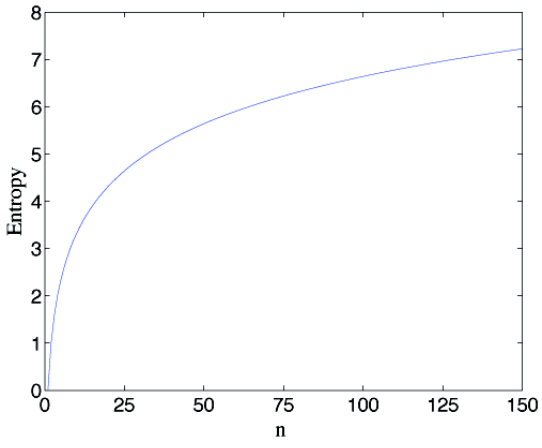
What about all three dice?

$$H = \log_2(4 \times 6 \times 20)$$
$$= \log_2(4) + \log_2(6) + \log_2(20)$$
$$= 8.9$$

H is Entropy = …

- The number of yes-no questions required to specify the state of the system
- If n is the number of equally likely states of the system:

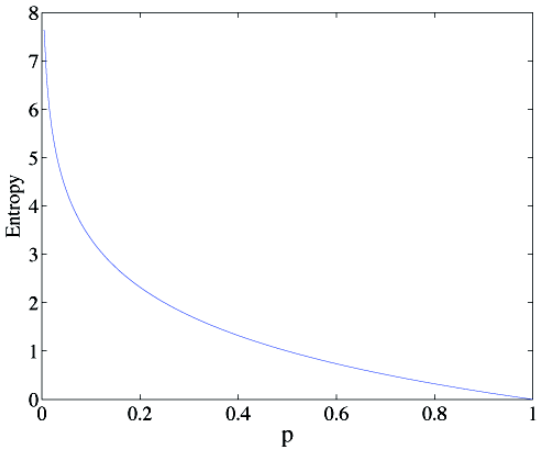
$$H = \log_2[n]$$



Rewrite the equations

$$H = \log_2[n] \qquad H = -\log_2\left[\frac{1}{n}\right]$$

$$H = -\log_2[p]$$



Non-Uniform Distributions

- When the probability of events isn't uniform, then more likely events convey less information
- Optimal code: $\left\lceil \log_2 \frac{1}{p(x)} \right\rceil$ for an event having probability $p(x)$
- The average number of bits needed to transmit a message
- Entropy: $H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$

Example 1: 8-sided die

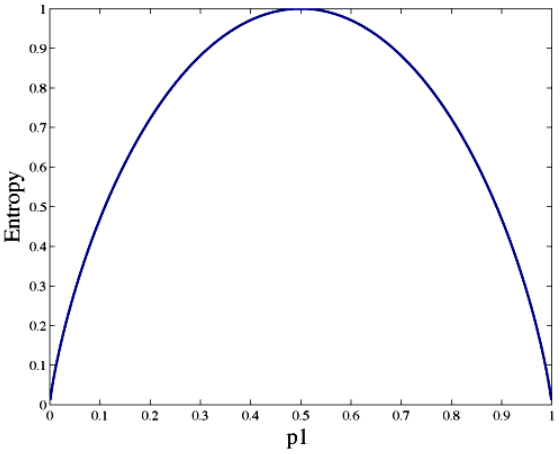
- Let x represent the result of rolling a (fair) 8-sided die.
- Entropy: $H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$

$$H(X) = \sum_{x \in X} \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} = \log_2 8 = 3$$

- The average length of the message required to transmit one of 8 equiprobable outcomes is 3 bits.
  - "1" "2" "3" "4" "5" "6" "7" "8"
    001 010 011 100 101 110 111 000

Entropy of a Weighted Coin

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$



- The more uncertain the result, the higher the entropy.
  - Fair coin: H(X) = 1.0
- The more certain the result, the lower the entropy.
  - Completely biased coin: H(X) = 0.0

Example 2: Simplified Polynesian

P	T	K	A	I	U
0,125	0,25	0,125	0,25	0,125	0,125

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

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

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

Example 2: Simplified Polynesian

Recall: $H = \log_2(6) = 2.585$ bits

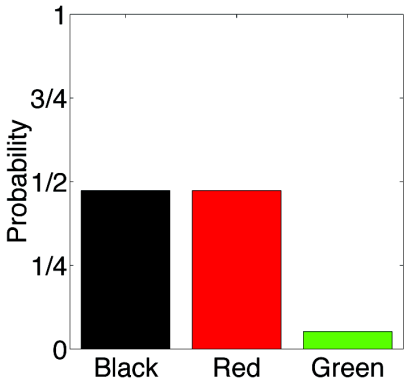
P	T	K	A	I	U
0,125	0,25	0,125	0,25	0,125	0,125

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

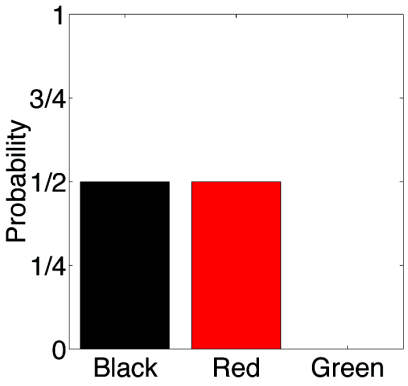
$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}]$$

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

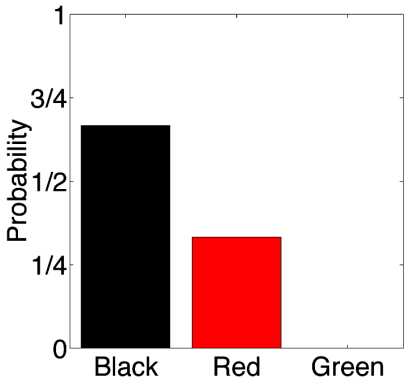
1.



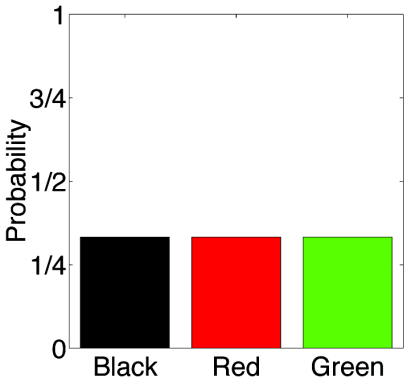
2.



3.



4.



Joint Entropy

- In language, the likelihood of events depends on the preceding outcomes (e.g. prior words).
- Joint entropy: the amount of information necessary to specify the value of two discrete random variables:

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

Polynesian revisited

- Assume the following (slightly different) per-letter frequencies:

p	t	k	a	i	u
1/16	3/8	1/16	1/4	1/8	1/8

  - H(X)= 2x1/16 $\log_2$ 16 + 2x1/8 $\log_2$ 8 + 1/4 $\log_2$ 4 + 3/8 $\log_2$ 8/3
    = 1/2+3/4+1/2+3/8 log2 8/3 = 2.28 per letter

  - Suppose we discover that, in Simplified Polynesian words consist of CV sequences. (margin probs are per syllable, not per letter)

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

- We can calculate H(C,V) from the table, i.e. treat each possible syllable as an event:

  - H(C,V) =1/4 log2 16 + 6/16 log2 16/3 + 3/8 log2 8/3
    = 2.436 per syllable (1.218 per letter)

Polynesian revisited

- Assume the following (slightly different) per-letter frequencies:

p	t	k	a	i	u
1/16	3/8	1/16	1/4	1/8	1/8

  - H(X)= 2x1/16 $\log_2$ 16 + 2x1/8 $\log_2$ 8 + 1/4 $\log_2$ 4 + 3/8 $\log_2$ 8/3
    = 1/2+3/4+1/2+3/8 log2 8/3 = 2.28 per letter

  - Suppose we discover that, in Simplified Polynesian words consist of CV sequences. (margin probs are per syllable, not per letter)

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

- We can calculate H(C,V) from the table, i.e. treat each possible syllable as an event:

  - H(C,V) =1/4 log2 16 + 6/16 log2 16/3 + 3/8 log2 8/3
    = 2.436 per syllable (1.218 per letter)

Conditional Entropy

- Conditional entropy: the amount of information needed to transmit Y, given that message X has been transmitted:

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x)$$

$$= \sum_{x \in X} p(x) \left[- \sum_{y \in Y} p(y \mid x) \log_2 (y \mid x) \right]$$

$$= - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(y \mid x)$$

Chain rule for joint entropy

- Chain rule for entropy:

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

$$= -E_{p(x,y)} (\log_2 p(x,y))$$

$$= -E_{p(x,y)} (\log_2 p(x) p(y \mid x))$$

$$= -E_{p(x,y)} (\log_2 p(x) + \log_2 (p(y \mid x))$$

$$= -E_{p(x)} (\log_2 p(x)) - E_{p(x,y)} (\log_2 p(y \mid x))$$

$$= H(X) + H(Y \mid X)$$

- In general: $H(X_1,...,X_n) = H(X_1) + H(X_2 \mid X_1) + ... + H(X_n \mid X_1,...,X_{n-1})$

H(C,V)=H(C)+H(V|C)

$$H(C) = 2 \times \frac{1}{8}\log_2 8 + \frac{3}{4}\log_2 \frac{4}{3}$$
$$= \frac{3}{4} + \frac{3}{4}(2 - \log_2 3) = \frac{9}{4} - \frac{3}{4}\log_2 3 \approx 1.061$$

$$H(V \mid C) = \sum_{c=p,t,k} p(C=c) H(V \mid C=c)$$
$$= \frac{1}{8}H(V \mid p) + \frac{1}{8}H(V \mid k) + \frac{3}{4}H(V \mid t)$$
$$= \frac{1}{8}H(\frac{1}{2},\frac{1}{2},0) + \frac{1}{8}H(\frac{1}{2},0,\frac{1}{2}) + \frac{3}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{4})$$
$$= 2 \times \frac{1}{8} \times 1 + \frac{3}{4}(\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2)$$
$$= \frac{1}{4} + \frac{3}{8} + \frac{3}{8} + \frac{3}{8} = \frac{11}{8} \approx 1.375$$

	p	t	k	
a	1/16	3/8	1/16	1/2
i	1/16	3/16	0	1/4
u	0	3/16	1/16	1/4
	1/8	3/4	1/8	

$$H(C,V) = H(C) + H(V \mid C)$$
$$= 1.061 + 1.375 = 2.436$$

Information …

- measures the uncertainty of a random variable
- indicates the number of bits of information encoded by that outcome
- is determined by the probability of the outcome,
  - which may be determined by context
- the more structure in the system, the lower the entropy, i.e. the average # of bits per event