An Overview of the Mathematical Theory of Communication Particularly for Philosophers Interested in Information

Simon D'Alfonso

The *Mathematical Theory of Communication* (or *Information Theory* as it is also known as) was developed primarily by Claude Shannon in the 1940s [10]. It measures the information (structured data) generated by an event (outcome) in terms of the event's statistical probability and is concerned with the transmission of such structured data over (noisy) communication channels.

The Shannon/Weaver communication model with which MTC is concerned is given in Figure 1.[1]

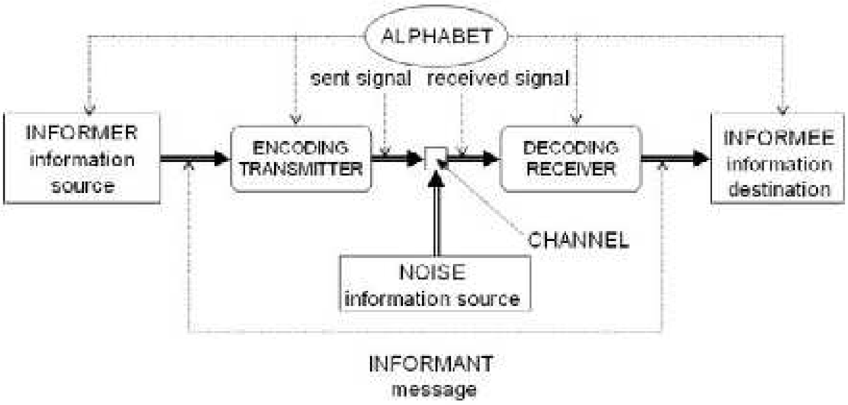

Figure 1: Shannon Weaver communication model ([6])

A good example of this model in action is Internet telephony. John says 'Hello Sally' in starting a conversation with Sally over Skype. John is the informer or information source and the words he utters constitute the message. His computer receives this message via its microphone and digitally encodes it in preparation for transmission. The encoding is done in a binary alphabet, consisting conceptually of '0' and '1's. The signal for this encoded message is sent over the Internet, which is the communication channel. Along the way some noise is added to the message, which interferes with the data corresponding to 'Sally'. The received signal is decoded by Sally's computer, converted into audio and played through the speakers. Sally, the informee at the information destination, hears 'Hello Sal**', where * stands for unintelligible crackles due to the noise in the decoded signal.[2]

[1] The applicability of this fundamental general communication model goes beyond MTC and it underlies many accounts of information and its transmission.

[2] Despite the simplicity of this high-level account, it adequately illustrates the fundamental components involved. Beyond this account there are richer ones to be given. Firstly, each process can be explained in greater degrees of detail. Secondly, this communication model can apply to other processes in the whole picture; for example, the communication involved in the transmission of sound from Sally's ear to a signal in her brain.

In order to successfully carry out such communication, there are several factors that need to be worked out. What is the (minimum) amount of information required for the message and how can it be encoded? How can unwanted equivocation and noise in the communication channel be dealt with? What is the channel's capacity and how does this determine the ultimate rate of data transmission? Since MTC addresses these questions it plays a central role in achieving the execution of this model. Here is a brief outline of the basic mathematical ideas behind MTC.[3]

Let $\mathbb{S}$ stand for some event/outcome/source which generates/emits symbols in some alphabet $\mathbb{A}$ which consists of n symbols. As three examples of this template, consider the following:

1. $\mathbb{S}$ is the tossing of a coin. $\mathbb{A}$ consists of two symbols, 'heads' and 'tails'.
2. $\mathbb{S}$ is the rolling of a die. $\mathbb{A}$ consists of six symbols, the numbers 1-6.
3. $\mathbb{S}$ is the drawing of a name in an eight-person raffle. $\mathbb{A}$ consists of each of the eight participants names.

The information measure associated with an event is proportional to the amount of certainty it reduces. For an event where all symbols have an equal probability of occurring, the probability of any one event occurring is $\frac{1}{n}$. The greater n is to begin with, the greater the number of initial possibilities, therefore the greater the reduction in uncertainty or data deficit. This is made mathematically precise with the following formulation. Given an alphabet of n equiprobable symbols, the information measure or entropy of the source is calculated with the following:

$$\log_2(n) \text{ bits}$$

Going back to the above three examples:

1. The outcome of a coin toss generates $\log_2(2) = 1$ bit of information
2. The outcome of a die roll generates $\log_2(6) = 2.58$ bits of information
3. The outcome of an eight-person raffle generates $\log_2(8) = 3$ bits of information

In cases where there is only one possible outcome uncertainty is zero and thus so is information. In terms of card types, the random selection of a card from a standard 52-card deck generates $\log_2(52) = 5.7$ bits of information. But the selection of a card from a deck consisting of 52 cards, all king of spades, generates $\log_2(1) = 0$ bits of information.

Skipping over the technical details and derivations, the general formula for the entropy (H) of a source, the average quantity of information it produces (in bits per symbol), is given by

$$H = -\sum_{i=1}^{n} \mathrm{Pr}(i)\log_2\mathrm{Pr}(i) \text{ bits per symbol} \tag{1}$$

[3] For a very accessible introduction to MTC, see [8]. For a considerably in-depth textbook see [3]. Shannon's original paper is [10].

for each of the n possible outcomes/symbols i. When all outcomes are equiprobable, this equation reduces to the formula above. When the source's outcomes are not all equiprobable things become more interesting.

Let us start with a fair coin, so that Pr('heads') = Pr('tails') = 0.5. Plugging these figures into Equation 1, we get:

$H = -(\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2})) = -(\frac{1}{2} \times -1 + \frac{1}{2} \times -1) = 1$ bit

which is the same as $\log_2(2) = 1$ bit.

But now consider a biased coin, such that Pr('heads') = 0.3 and Pr('tails') = 0.7. Plugging these figures into equation 1, we get:

$H = -(0.3 \times \log_2(0.3) + 0.7 \times \log_2(0.)) = -((0.3 \times -1.737) + (0.7 \times -0.515)) = 0.8816$ bits

So the biased coin generates less information than the fair coin. This is because the overall uncertainty in the biased coin case is less than in the fair coin case; with the former case there is a higher chance of 'tails' and lower change of 'heads' so in a sense any outcome is less surprising. This is all mathematically determined by the structure of the formula for H. The occurrence of some particular symbol generates some amount of information; the lower the probability of it occurring the higher the information generated. This is represented with the $\log_2 \mathrm{Pr}(i)$ part. Although a lower probability means more information on an individual basis, with the average calculation this is regulated and diminished with the multiplication by its own probability. This balance is why H takes its highest value when all of a source's potential symbols are equiprobable.

Equation 1 represents a fundamental limit. It represents the lower limit on the expected number of symbols ('0's and '1's)[4] required to devise a coding scheme for the outcomes of an event, irrespective of the coding method employed. It represents the most efficient way that the signals for an event can be encoded. It is in this sense that H is the unique measure of information quantity.

This point can be appreciated with the simplest of examples. Take the tossing of two fair coins (h = heads, t = tails). John is to toss the coins and communicate the outcome to Sally by sending her a binary digital message. Since the coins are fair, $\mathrm{Pr}(h,h) = \mathrm{Pr}(h,t) = \mathrm{Pr}(t,h) = \mathrm{Pr}(t,t) = 0.25$. The number of bits required to code for the tossing of these coins is two ($H = 2$); it is simply not possible on average to encode this information in less than two bits. Given this, John and Sally agree on the following encoding scheme:

- $(h,h) = 00$
- $(h,t) = 01$
- $(t,h) = 10$
- $(t,t) = 11$

As an example, the string which encodes the four outcomes (h,h), (t,t), (h,t) and (h,h) is '00110100'.

[4] This is the standard set of symbols for a binary alphabet ($n = 2$). n is the same as the base of the logarithm in H, which does not have to be two. A base that has also been used is ten (in this case the unit of information is called a 'Hartley', in honour of Ralph Hartley, who originated the measure). Natural logarithms have also been used (in this case the unit of information is called a 'nat').

Now, modify this coin scenario so that the outcomes have the following probabilities:

- $\Pr(h, h) = 0.5$
- $\Pr(h, t) = 0.25$
- $\Pr(t, h) = 0.125$
- $\Pr(t, t) = 0.125$

Given this probability distribution, $H = 1.75$ bits. As we have just seen, this means that the lower limit on the average number of symbols required to code each tossing of the two coins is 1.75. How would such an encoding go? The basic idea is to assign fewer bits to the encoding of more probable outcomes and more bits to the encoding of less probable outcomes. Since (h, h) is the most probable outcome, fewer bits should be used to encode it. This way, the number of expected bits required is minimised.

The most efficient coding to capture this connection between higher probability of an outcome and more economical representation is:

- $(h, h) = 0$
- $(h, t) = 10$
- $(t, h) = 110$
- $(t, t) = 111$

If we treat the number of bits for each outcome as the information associated with that outcome, then we can plug these figures into the following formula and also get a calculation of 1.75:

$(0.5 \times 1) + (0.25 \times 2) + (0.125 \times 3) + (0.125 \times 3) = 1.75$ bits

In comparison to the previous example, the string which represents the four outcomes (h, h), (t, t), (h, t) and (h, h) using this encoding scheme is the shorter '0111100'. This optimal encoding scheme is the same as that which results from the Shannon-Fano coding method.[5]

The discussion of MTC thus far has involved fairly simple examples without any of the complications and complexities typically involved in realistic communication. To begin with, it has only considered the information source of perfect communication channels, where data is received if and only if it is sent. In real conditions, communications channels are subject to *equivocation* and *noise*. The former is data that is sent but never received and the latter is data that is received but not sent. The communication system as a whole involves both the possible outcomes that can originate from the information source $S = \{s_1, s_2, ...s_m\}$ and the possible signals that can be received at the information destination $R = \{r_1, r_2, ...r_n\}$. It is the statistical relations between S and R (the conditional probability that an element in one set occurs given the occurrence of an element from the other set) that determine the communication channel.

Here are a couple of examples taken from Dretske [4] that demonstrate how all this works.

[5] Whilst it produces an optimal encoding in this case, in general this method is suboptimal, in that it does not achieve the lowest possible expected code word length. Another method which does generally achieve the lowest possible expected code word length is Huffman coding (`http://en.wikipedia.org/wiki/Shannon-Fano_coding`).

Example A boss asks eight of his employees to select amongst themselves one individual to perform some task. Once this person is selected, they will inform the boss of their choice by writing the selected person's name down on a piece of paper and sending it to the boss. Each employer is uniquely named and each has an equal probability ($\frac{1}{8}$) of being selected using some random selection process. The information source that is the message generating selection process gives a reduction of 8 possibilities to 1 and it generates 3 bits of information; its entropy (H), the average amount of information associated with the source S is 3 bits, calculated as:

$$H = -\sum_{i=1}^{n} \mathrm{Pr}(i) \times \log_2 \mathrm{Pr}(i) = 3 \text{ bits}$$

for each of the n possible outcomes/symbols i.

There is no noise/equivocation in this communication system (each message can be traced to one outcome) so the amount of information received equals the amount of information generated.

□

Example A small modification is made to the previous example so that equivocation is greater than zero. Everything is the same except that for some reason, the employees decide that should Barbara be selected as a result of their selection process, they will write Herman's name down on the note instead. So the message 'Herman' would now be used if either Herman or Barbara were selected. This equivocation (that 'Herman' cannot be traced to one unique outcome) affects the information of the source, as the following calculations will demonstrate.

For each source event s_i and for some message r_a, the equivocation associated with r_a is calculated as:

$$E(r_a) = -\sum_{i=1}^{n} \mathrm{Pr}(s_i|r_a) \times \log_2 \mathrm{Pr}(s_i|r_a)$$

The average equivocation associated with a source is the probability-weighted sum of the equivocation associated with each message:

$$E = -\sum_{i=1}^{n} \mathrm{Pr}(r_i) \times E(r_i) \tag{2}$$

for each of the n messages r_i.

When the channel is perfect, equivocation is 0; either $\mathrm{Pr}(s_i|r_j)$ takes the value of 0 or $\log_2 \mathrm{Pr}(s_i|r_j)$ takes the value of 0.

If the channel is completely random, then $\mathrm{Pr}(s_i|r_j) = \frac{1}{8}$ for each message and equivocation is maximum (i.e. the equivocation equals the amount generated from the source), so no information is carried. In this scenario, the calculation would be:

$$
\begin{array}{rcl}
E & = & (8 \times \frac{1}{8} \times -(8 \times \frac{1}{8} \times \log_2(\frac{1}{8}))) \\
E & = & (1 \times -(8 \times \frac{1}{8} \times -3)) \\
E & = & (1 \times -(1 \times -3)) \\
E & = & 3
\end{array}
$$

In our current example, equivocation is somewhere between the maximum 3 and minimum 0. Since the message 'Herman' is responsible for the non-zero equivocation (all other messages have zero equivocation), we only need to calculate E() for when r is 'Herman' (r_H). This will involve the outcomes 'Herman' (H) and 'Barbara' (B)

$$
\begin{array}{rcl}
\mathrm{E}(r_H) & = & -[\Pr(\mathrm{B}|r_H)\log_2\Pr(\mathrm{B}|r_H) + \Pr(\mathrm{H}|r_H)\log_2\Pr(\mathrm{H}|r_H)] \\
\mathrm{E}(r_H) & = & -(\frac{1}{2} \times -1) + (\frac{1}{2} \times -1)) \\
\mathrm{E}(r_H) & = & 1
\end{array}
$$

Since r_H will appear if either H or B, $\Pr(r_H) = \frac{2}{8}$

Plugging this into Equation 2, we get

$$
\mathrm{E} \quad = \quad \frac{1}{4} \times 1 = 0.25
$$

Thus the average equivocation on the channel rises from 0 to 0.25 and the average amount of transmitted information is now 2.75.

□

Redundancy refers to the difference between the number of bits used to transmit a message and the number of bits of actual information in the message. Whilst redundancy minimisation through data compression is desirable, redundancy can also be a good thing, as it is used to deal with noise and equivocation. As the simplest of examples, if John says 'hello hello' to Sally, the second hello is redundant. But if the first hello becomes unintelligible due to noise/equivocation, then an intelligible second hello will serve to counter the noise/equivocation and communicate the information of original message. In technical digital communication, sophisticated error detection and correction algorithms economically use desired redundancy.

Another factor to briefly mention is that the probability distribution of the source can be conditional. In our examples, the probability distributions were fixed and the probability of one outcome was independent of any preceding outcome. The term for such a system is *ergodic.* Many realistic systems are non-ergodic. For example, you are about to be sent an English message character by character. At the start there is a probability distribution across the range of symbols (i.e. English alphabet characters). If an 'h' occurs as the first character in the message then the probability distribution changes. For example, the probability that the next character is a vowel would increase and the probabilities for 'h' and 'k' would decrease, effectively to zero, since there are no valid constructions in the English language with 'hh' or 'hk'. Whilst such complications and complexities are covered by MTC, the details are unnecessary for our purposes and need not detain us.

Once a message is encoded it can be transmitted through a communication channel.[6] Shannon came up with the following two theorems concerning information transmission rates over communication channels. Let C stand for the transmission rate of a channel, measured in bits per second

[6] Research into the practical concerns of communication was a key factor for Shannon, whose interest resulted from his work at AT&T Bell Labs. As a telephone company, they wanted to know what minimum capacities their networks needed in order to efficiently handle the amounts of traffic (data) they were expecting to deal with.

(bps). Shannon's theorem for noiseless channels:

> Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate of $C/H - \epsilon$ symbols per second over the channel where ϵ is arbitrarily small. It is not possible to transmit at an average rate greater than C/H.[9, p. 59.]

To deal with the presence of noise in practical applications, there is the corresponding theorem for a discrete channel with noise:

> Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \epsilon$ where ϵ is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$.[9, p. 71.]

As can be gathered, MTC covers several properties associated with an intuitive conception of information:

- information is quantifiable
- information quantity is inversely related to probability
- information can be encoded
- information is non-negative
- information is additive

Ultimately however MTC is a syntactic treatment of information that is not really concerned with semantic aspects. Although it deals with structured data that is potentially meaningful, any such meaning has no bearing on MTC's domain. Theories of semantic information on the other hand deal with data that is meaningful and the use of such data by semantic agents. The following examples serve to illustrate these points:

1. For MTC information is generated when one symbol is selected from a set of potential symbols. As we have seen, entropy, the measure of information associated with a source, is inversely related to probability. Given the English alphabet as a set of symbols, the string 'xjk' is less probable than the string 'dog', therefore according to MTC it has a higher entropy/information measure. Despite this, the word 'dog', unlike 'xjk', is meaningful to an English informee and potentially informative. So in this sense, with MTC gibberish yields more information than probable yet meaningful strings.
2. Take a network over which statements in the English language are encoded into ASCII [7] messages and transmitted. The encoding of each character requires 7 bits. Now, consider the following three strings:

[7] American Standard Code for Information Interchange: `http://en.wikipedia.org/wiki/ASCII`

   - the an two green four cat !?down downx
   - Colourless green ideas sleep furiously
   - The native grass grew nicely in spring

   Although each message uses the same number of bits ($7 \times 38 = 266$), the first is not well-formed in accordance with the syntax of the English language and the second is well-formed but is not meaningful. Only the third is well-formed and meaningful and hence can be considered to be semantic information.

3. Consider a basic propositional logic framework. Say that for each symbol in a statement 1 unit of data is required in its encoded message. Consider the following three strings:

   - $A\neg B$
   - $A \vee B$
   - $A \wedge B$

   Each of these statements contains the same quantity of syntactic data. The first however, is not well-formed. Whilst the second and third are well-formed, according to the meanings of the connectives $\vee$ and $\wedge$, there is a sense in which $A \wedge B$ is more informative than $A \vee B$.

So MTC is ultimately about the quantification and communication of syntactic information or data. It

> approaches information as physical phenomenon, syntactically. It is interested not in the usefulness, relevance, interpretation, or aboutness of data but in the level of detail and frequency in the uninterpreted data (signals or messages). It provides a successful mathematical theory because its central question is whether and how much data, not what information is conveyed.[5, p. 561.]

Whilst it is not meant to deal with the semantic aspects of information, since MTC deals with the data that constitutes semantic information it is still relevant and provides scientific constraints for theories of semantic information and a philosophy of information [7]. Ideas from MTC have actually found application in methods of quantitatively measuring semantic information in communication [1]. Furthermore, MTC can serve as a starting point and guide for a semantical theory of information [8]. Chapman [2] argues for a stronger link between Shannon information and semantic information.

References

[1] Jie. Bao. Towards a theory of semantic communication. Technical report, Rensselaer Polytechnic Institute, 2011. URL = <http://www.cs.rpi.edu/~baojie/pub/2011-03-28_nsw_tr.pdf>.

[2] David Chapman. Information, meaning and context. In Magnus Ramage and David Chapman, editors, *Perspectives on Information*, pages 36–50. Routledge, 2011.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, second edition, 2006.

[4] Fred Dretske. *Knowledge and the Flow of Information.* MIT Press, Cambridge, MA, 1981.

[8] See [4] for example

[5] L. Floridi. Open problems in the philosophy of information. *Metaphilosophy*, 35(4):554–582, 2004.

[6] Luciano Floridi. Semantic conceptions of information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, summer 2009 edition, 2009. URL = <http://plato.stanford.edu/archives/sum2009/entries/information-semantic/>.

[7] Luciano Floridi. The philosophy of information as a conceptual framework. *Knowledge, Technology and Policy*, 23(1-2):253–281, 2010.

[8] John R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise.* Dover Publications, New York, second edition, 1980.

[9] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, 1949.

[10] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. URL = <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.