Formal Models of Language: Language as Information

Paula Buttery

Easter 2020

Languages are mutually understandable communication systems, in which information is transmitted from speaker to hearer. Information Theory provides a mathematical basis on which to discuss transmission of information across (noisy) channels. A question then is whether information theory can inform theories of human communication and language.

You cover Information Theory fundamentals in the *Information Theory* course. I will not repeat all those details here but rather summarise the concepts we need to know to explore some natural language theories.

1. Information Theory Basics and Natural Language

Information ***sources*** produce ***information*** as ***events*** or ***messages***. These may be represented by a random variable X over a discrete set of symbols (or alphabet) $\mathcal{X}$.[1]

Entropy (or ***self-information***) may be thought of as:

- the average amount of information produced by a source;
- the average amount of uncertainty of a random variable;
- the average amount of information we gain when receiving a message from a source;[2]
- the average amount of information we lack before receiving the message;
- the average amount of uncertainty we have in a message we are about to receive.

Entropy is measured in ***bits***. If a source produces M messages with equal frequency then the entropy (the information produced by the source per message) is $H = log_2 M$.[3] Entropy provides a lower limit on the number of bits we need to represent an event space.[4]

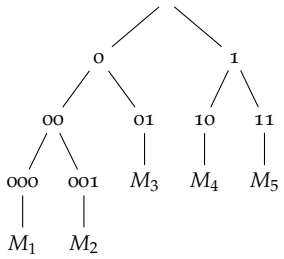
Entropy as described so far assumes that all messages are equally likely to occur. This will not work well for describing the information gained when receiving successive words of natural language: intuitively, it seems like the information gained on receiving the highly frequent word *the* should be less than the information gained on receiving the infrequent word *yak*.[5] Similarly, if we considered the information transmitted by successive letters in a word: it seems like the vowels are providing less information than the consonants.[6]

[1] For example, for a dice roll $\mathcal{X} = \{1,2,3,4,5,6\}$; for a source that produces characters of written English $\mathcal{X} = \{a...z,\}$. Note that, when discussing natural languages, a message might be a letter of the language's *alphabet* or a *word* (depending on the level of linguistic enquiry).

[2] Thinking of entropy as information received is useful when we consider a language to be a communication system, in which information is transmitted from speaker to hearer.

[3] Notice that $log_2 1 = 0$: i.e. when there is only one possible message that we can receive, it conveys no information and no uncertainty.

[4] You will cover this in detail in *Information Theory*, but entropy also gives us a lower limit on the average number of bits you need per message code. If a source can produce M messages then we need M codes to represent all the possible messages. For example, for a source that can produce 5 messages:



$$\begin{aligned} \textit{average length} &= \frac{(3*2)+(2*3)}{5} = 2.4 \\ &\approx H(5) = log_2 5 = 2.32 \end{aligned}$$

[5] Remember that we know that words in a text exhibit a power-law probability distribution (Zipf's law).

[6] Indeed, Hebrew is often written without vowels (called unpointed).

When the probability of events isn't uniform, then more likely events convey less information. ***Surprisal*** is a measure that allows us to calculate the information of non-uniform events in bits.

entropy and surprisal

If the probability of a message, x, is $p(x)$ then the surprisal of x is:

$$s(x) = \log_2\left(\frac{1}{p(x)}\right) = -\log_2 p(x)$$

Surprisal is also measured in ***bits*** and gives us a measure of information that is inversely proportional to the probability of a message occurring (i.e probable events convey a small amount of information and improbable events a large amount of information). From an encoding point of view, surprisal provides an indication of the number of bits we would want to assign a message's code. It is efficient to give probable items (with low surprisal) a small bit code because we have to transmit them often. The average information (entropy) produced per message by a non-uniform message source[7] is the weighted sum of the surprisal (the average surprise): [8]

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

When all messages are assumed to be equally likely we call that a ***oth order model*** of the source. The weighted sum surprisal (which takes account of the probability of the messages) is a ***1st order model***. The 1st order model assumes independence of the messages. A ***2nd-order model*** takes account of context. That is, it models the probability of receiving message y given that we have just received message x (i.e. $p(y|x)$).[9]

Conditional entropy, $H(Y|X)$, is the average amount of information needed to transmit random variable Y, given that random variable X has been transmitted:

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

Joint entropy, $H(X,Y)$, is the amount of information needed on average to specify two discrete random variables:

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

The ***Chain rule*** connects joint and conditional entropy:

$$H(X,Y) = H(X) + H(Y|X)$$

$$H(X_1...X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1...X_{n-1})$$

Mutual information, $I(X;Y)$, indicates the information Y contains about X. It is a measure of the reduction in uncertainty of one random variable due to knowing about another. You can also think of $I(X;Y)$ as being the amount of information one random variable contains about another.

[7] or, if you like, the average bit code length per message

[8] Note, that when all M items in $\mathcal{X}$ are equally likely (i.e. $p(x) = \frac{1}{M}$) then $H(X) = -\log_2 p(x) = \log_2 M$ as on previous page.

[9] In natural language, we often talk about ***n-gram language models***. A ***bi-gram*** model takes account of the previous item when predicting the next item. In general an ***n-gram*** model takes account of the previous $n-1$ items.

conditional entropy, joint entropy, mutual information

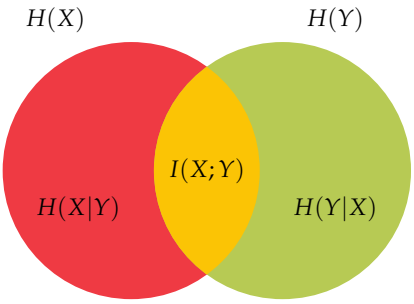


Figure 1: The interaction between conditional entropy, joint entropy and mutual information.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Relative entropy,[10] $D(p \parallel q)$, can be calculated for two probability mass functions over the same event space. It is a measure of how different the two probability distributions are[11] (and can be thought of as the average number of bits wasted by encoding events with distribution p using a code based on distribution q). For $p(x)$ and $q(x)$ the relative entropy is given by:

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{p(x)}{q(x)}\right) =$$

[10] also called Kullback-Leibler divergence.

[11] Mutual information turns out to be a measure of how far the joint distribution is from independence

$$I(X;Y) = D(p(x,y) \parallel p(x)p(y))$$

relative entropy, cross entropy

If a random variable X has a true probability distribution $p(x)$ but is modelled by $q(x)$ then the ***cross entropy***,[12] is given by:

$$\begin{aligned} H(X,q) &= H(X) + D(p \parallel q) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) + \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{p(x)}{q(x)}\right) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log_2 q(x) \end{aligned}$$

[12] Cross-Entropy is useful for measuring the performance of a classification model which outputs a probability distribution over the classes. Cross-entropy increases as the predicted probabilities diverge from the actual labels.

For a stochastic process we can calculate ***entropy rate***, H_{rate}. The entropy of a language is the limit of the entropy rate of a sample of the language, as the sample gets longer and longer:

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1 ... X_n)$$

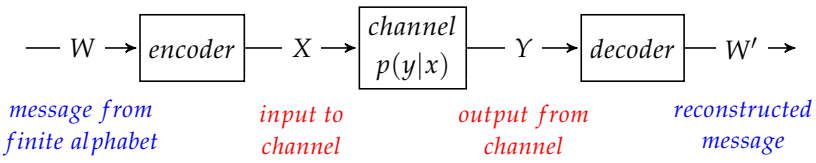
entropy rate

A quick reminder about ***maximum likelihood estimation***: often we do not know the real probability distribution associated with a random variable but we can estimate it using frequency counts. For instance the probability of a word sequence $w_1 ... w_n$ might be estimated as follows:

$$p_{mle}(w_1 ... w_n) = \frac{count(w_1 ... w_n)}{count(w_1 ... w_{n-1})}$$

2. Noisy Channel Theorem

A communication channel is any means by which a message can be conveyed; any noise in the channel establishes the relationship between input and output.



noisy channel

The input to the channel has an entropy $H(X)$ and the output has an entropy of $H(Y)$. An input-output pair has a probability $p(x,y) = p(x)p(y|x)$ and the joint entropy of all pairs is $H(X,Y)$ (which can be no larger than the sum of the individual entropies). When X and Y are totally independent they share no mutual information and then $H(X,Y) = H(X) + H(Y)$.

The ***capacity*** of a channel is the maximum of the mutual information of X and Y over all input distributions of the input $p(X)$:

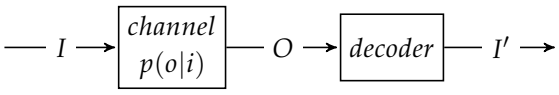
channel capacity

$$C = \max_{p(X)} I(X;Y)$$

Channel capacity, C, is the rate we can transmit information through a channel with an arbitrarily low probability of not being able to recover the input from the output. As long as transmission rate is less than C we don't need to worry about errors (the optimal transmission rate being C).[13]

[13] Ideally a message will be transmitted across a channel as efficiently as possible while retaining enough redundancy for errors to be detected and corrected. In practical applications we reach the channel capacity by designing an encoding for the input that maximises mutual information.

The noisy channel has inspired a Natural Language Processing (NLP) problem-solving ***framework***. In this framework we assume there is no control over the language encoding and start at the point where the already encoded input enters the channel:



Many problems can be thought of as trying to find the most likely input given an output from the channel:[14]

$$I' = \operatorname*{argmax}_i p(i|o) = \operatorname*{argmax}_i p(i)p(o|i)$$

[14] Examples include speech recognition, machine translation, spelling correction, character recognition, ...

where $p(i)$ is the probability of the input (usually an n-gram language model); and $p(o|i)$ is the ***channel probability*** (i.e. a model of the probability of getting an output from the channel given the input).[15]

[15] We can use Bayes' theorem to estimate $p(i|o)$ (which is often difficult to estimate directly and reliably):

$$p(i|o) = \frac{p(o|i)p(i)}{p(o)}$$

Note that $p(o)$ will have no effect on *argmax* function.

Information theory and natural language theories

Information theoretic approaches are being used to help us consider the following questions about natural language:

- Is natural language a good code from an information-theoretic perspective? Highly efficient codes make use of regularities in the messages from the source using shorter codes for more probable messages. Does this turn out to be true for natural language? [Piantadosi et al., 2011]
- Where speakers have a choice between several variants to encode their message, which variant will they prefer? Are constant rates of transmission preferred for lexical unit? Within the bounds defined by grammar, do speakers prefer utterances that distribute information uniformly across the signal? [Genzel and Charniak, 2002, Aylett and Turk, 2004, Jaeger, 2010]
- Can the noisy channel account for typological variations in the world's languages? If word order provides context that is informative about meaning does this account for observed word order in the world's languages? [Gibson et al., 2013]

- Why has natural language evolved to be so ambiguous? Does ambiguity have any communicative benefit? Is it the case that efficient communication systems will necessarily be globally ambiguous when context is informative? [Piantadosi et al., 2012]

References

Matthew Aylett and Alice Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56, 2004. DOI: 10.1177/00238309040470010201. URL https://doi.org/10.1177/00238309040470010201. PMID: 15298329.

Dmitriy Genzel and Eugene Charniak. Entropy rate constancy in text. In *40th Annual Meeting of Association for Computational Linguistics*, Philidelphia, July 2002. Association for Computational Linguistics.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088, 2013. DOI: 10.1177/0956797612463705. URL https://doi.org/10.1177/0956797612463705. PMID: 23649563.

T. Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23 – 62, 2010. ISSN 0010-0285. DOI: https://doi.org/10.1016/j.cogpsych.2010.02.002. URL http://www.sciencedirect.com/science/article/pii/S0010028510000083.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011. ISSN 0027-8424. DOI: 10.1073/pnas.1012551108. URL http://www.pnas.org/content/108/9/3526.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280 – 291, 2012. ISSN 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2011.10.004. URL http://www.sciencedirect.com/science/article/pii/S0010027711002496.