

Entropy of Human Language

One of the most important sources of information (as for everyday life and also application of data compression) is written and spoken human language. This is apparently not a stationary ergodic process, so the analysis is not straightforward. We shall here try to find the entropy of English. (Different languages have different entropies, but the tools for analysing them are about the same.) Note that the entropy of a source also plays a central role in *cryptanalysis*.

We start by looking at various stochastic approximations to English. We assume that the alphabet consists of 26 letters and the space symbol (ignoring upper and lower case and punctuation).

© Patric Östergård

Approximation of English Text (2)

From Shannon's "A Mathematical Theory of Communication":

Zero-order approximation: (independent and equiprobable symbols)
XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD
QPAAMKBZAACIBZLHJQD

First-order approximation: (letter frequency matches that of English text)
OCRO HLI RGWR NMIELWIS EU L NBNESBYA TH EEI
ALHENHTTPA OOBTTVA NAH BRL

Second-order approximation: (frequency of letter pairs matches that of English text)
ON IE ANTSOUTINYS ARE T INCTORE ST BE
S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE
FUSO TIZIN ANDY TOBE SEACE CTISBE

© Patric Östergård

Approximation of English Text (1)

The frequencies of letters is not uniform, but range from 13% (E) to 0.1% (Q,Z). (But the novel *Gadsby* by E. V. Wright does not contain a single E. . . .)

The frequencies of pairs of letter is clearly not uniform either: Q is always followed by U and the most frequent pair is TH with a frequency of 3.7%.

Proceeding in this way, higher order conditional probabilities can be estimated and more complex models built. However, to make estimation easier, a sample of text (from books, for example) is preferred.

© Patric Östergård

Approximation of English Text (3)

First-order word model: (words independent with frequency matching that of English text)
REPRESENTING AND SPEEDILY I
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.

Second-order word model: (frequency of word pairs matches that of English text)
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

© Patric Östergård

Approximation of English Text (4)

The entropies of the previously listed approximations get closer and closer to the entropy of English.

Zero-order approximation: entropy is $\log 27 = 4.76$ bits per letter.

Second-order approximation: entropy is 4.03 bits per letter.

⋮

4th-order approximation: (omitted) entropy is 2.8 bits per letter.

⋮

© Patric Östergård

The Shannon Guessing Game (2)

Assuming that the human subject has an optimal strategy, any one knowing the optimal strategy can reconstruct the original text. Hence, the amount of information (and therefore the entropy) are the same in the text and in the sequence of guess numbers.

The guess numbers are not independent, but we can get an upper bound on the entropy by assuming that they are independent. Then the entropy is easily calculated from the distribution of guess numbers.

When Shannon carried out this experiment, he obtained a value of 1.3 bits per symbol for the entropy of English.

© Patric Östergård

The Shannon Guessing Game (1)

There are two methods, both related to gambling, that can be used to estimate the entropy of English. The first one is called the **Shannon guessing game**.

In the Shannon guessing game, a human subject is given a sample of text and asked to guess the next letter. The most probable letter is guessed first, then the second most probable letter, etc. The number of guesses required is recorded. One possible sequence is, for example, 12, 9, 4, 5, 1, 3, 4, 2, ...

© Patric Östergård

Gambling Estimate (1)

In a **gambling estimate**, a human subject gambles on the next letter in a sample. The payoff is 27-for-1 on the correct letter (pay 1, receive 27 if correct).

Sequential betting is equivalent to betting on the entire sequence, so the payoff after n letters can be written as

$$S_n = (27)^n b(X_1, X_2, \dots, X_n),$$

where $b(X_1, X_2, \dots, X_n)$ is the fraction of the gambler's wealth bet on this sequence.

© Patric Östergård

Gambling Estimate (2)

After n rounds of betting, the expected log wealth satisfies

$$\begin{aligned}
 E \frac{1}{n} \log S_n &= \log 27 + \frac{1}{n} E \log b(X_1, X_2, \dots, X_n) \\
 &= \log 27 + \frac{1}{n} \sum_{x^n} p(x^n) \log b(x^n) \\
 &= \log 27 - \frac{1}{n} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{b(x^n)} + \frac{1}{n} \sum_{x^n} p(x^n) \log p(x^n) \\
 &= \log 27 - \frac{1}{n} D(p(x^n) \| b(x^n)) - \frac{1}{n} H(X_1, X_2, \dots, X_n) \\
 &\leq \log 27 - \frac{1}{n} H(X_1, X_2, \dots, X_n) \leq \log 27 - H(\mathcal{X}).
 \end{aligned}$$

©Patric Östergård

Performance of Available Software

A comparison of compression of English text (and other sources) can be found at <http://compression.ca/>.

The best compression of the sources *The Three Musketeers*, *Anne of Green Gables*, and the *1995 CIA World Fact Book* is (in November 2002) 1.30 bits/byte (ASCII text has one byte per letter) by the program ENTROPY 0.5, specialized on compressing English text. The test files obviously contain punctuation and upper and lower case letters. The (universal coding) program `gzip` achieves 2.56 bits/byte in the test.

©Patric Östergård

Gambling Estimate (3)

As $E \frac{1}{n} \log S_n \leq \log 27 - H(\mathcal{X})$, we have

$H(\mathcal{X}) \leq \log 27 - E \frac{1}{n} \log S_n = \hat{H}$. The upper bound estimate \hat{H} converges to H with probability one if English is ergodic and the gambler uses $b(x^n) = p(x^n)$.

Experiments with this approach has lead to an estimate of 1.34 bits per letter for the entropy of English.

©Patric Östergård

The Second Law of Thermodynamics (1)

One of the basic laws of physics is the **second law of thermodynamics**: *The entropy of an isolated system is non-decreasing*. What is the relationship between this law and the entropy function of information theory?

In statistical thermodynamics, entropy is often defined as the logarithm of the number of microstates in the system \Rightarrow This corresponds to our notation of entropy if all the states are equally likely.

©Patric Östergård

The Second Law of Thermodynamics (2)

We model the isolated system as a Markov chain. Obviously, knowing the present state, the future of the system is independent of the past. There are now several independent interpretations of the second law. According to these

- entropy does not always increase, but
- *relative* entropy always decreases.

In the late 1940s, John von Neumann, a pioneer of the computer age, is said to have advised communication-theorist Claude E. Shannon to start using the term entropy when discussing information because "no one knows what entropy really is, so in a debate you will always have the advantage."

©Patric Östergård

The Second Law of Thermodynamics (4)

Since p and q are derived from the Markov chain, we have $p(x_{n+1}|x_n) = q(x_{n+1}|x_n) = r(x_{n+1}|x_n)$ and hence $D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = 0$. Thus

$$D(p(x_{n+1})||q(x_{n+1})) \leq D(p(x_n)||q(x_n)).$$

2. *Relative entropy* $D(\mu_n||\mu)$ *decreases with* n . (A stationary distribution is denoted by μ .) To see this, insert $q(x_n) = \mu'_n = \mu$ in the final formula above, and since μ is a stationary distribution, $\mu'_{n+1} = \mu$, and we get

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu).$$

©Patric Östergård

The Second Law of Thermodynamics (3)

1. *Relative entropy* $D(\mu_n||\mu'_n)$ *decreases with* n . Let μ_n and μ'_n be two probability distributions of a Markov chain at time n , with the corresponding joint mass function denoted by p and q . The probability transition function for the Markov chain is denoted by $r(\cdot|\cdot)$. By the chain rule for relative entropy, we get that $D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) =$

$$\begin{aligned} D(p(x_n)||q(x_n)) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = \\ D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})). \end{aligned}$$

©Patric Östergård

The Second Law of Thermodynamics (5)

3. *Entropy increases if the stationary distribution is uniform.* In general, the fact that the relative entropy decreases does not imply that the entropy increases. (Example: A Markov chain with a non-uniform stationary distribution, starting from the uniform distribution: entropy decreases.) If, however, the stationary distribution is uniform, we get

$$D(\mu_n||\mu) = \log |\mathcal{X}| - H(\mu_n) = \log |\mathcal{X}| - H(X_n).$$

As the relative entropy decreases, the entropy increases.

©Patric Östergård

The Second Law of Thermodynamics (6)

4. *The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process.* With a stationary Markov process, $H(X_n)$ is constant. For the conditional entropy, $H(X_n|X_1)$, on the other hand, we get

$$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1, X_2) \\ &= H(X_n|X_2) \\ &= H(X_{n-1}|X_1). \end{aligned}$$

©Patric Östergård

The Second Law of Thermodynamics (7)

5. *Shuffling increases entropy.* If T is a shuffle (permutation) of a deck of cards, if X is the initial (random) position of the cards in the deck, and if X and T are independent, we have

$$\begin{aligned} H(TX) &\geq H(TX|T) \\ &= H(T^{-1}TX|T) \\ &= H(X|T) \\ &= H(X). \end{aligned}$$

©Patric Östergård