**MAKERERE UNIVERSITY**

**COLLEGE OF ENGINEERING, DESIGN, ART AND TECHNOLOGY**

**DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING**

**Communication Engineering II**

**Assignment 1**

**Year 4**

**NAME: KIRUMIRA Aziz**

**REG No: 09/U/2906/PSA**

**Student number: 209003274**

**Course: BSc. ELECTRICAL ENGINEERING**

# Introduction

In 1948, Shannon published his paper “A Mathematical Theory of Communication” in the Bell Systems Technical Journal. He showed how information could be quantified with absolute precision, and demonstrated the essential unity of all information media. Telephone signals, text, radio waves, and pictures, essentially every mode of communication, could be encoded in bits. The paper provided a “blueprint for the digital age”. In it, he presents language as an example of an information source which one might communicate. He provides an estimate of the entropy of English and gives an analysis of its redundancy as related to crossword puzzles.

Shannon published another paper on the subject in 1951 called “Prediction and Entropy of Printed English”. In this paper Shannon gives his full attention to the entropy of English. He gives more rigorous mathematical approximations of the entropy of English words and letters but the main result of the paper was his human based approximation method. Shannon ﬁrst derives some bounds for the entropy of an ideal information source predictor. Assuming a person is such a predictor, he conducts some experiments to estimate the prediction properties of humans. Using these results, he estimates entropy.

# Estimating the Entropy of English

Entropy is a measure of the uncertainty of a random variable. The entropy *H(X)* of a discrete random variable *X* is defined by

Where X is a discrete random variable with alphabet and probability mass function The log is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit.

Entropy, as defined by Shannon, is the uncertainty regarding which symbols are chosen from a set of symbols with given a priori probabilities. Since information is a decrease in uncertainty, we may regard entropy as the information required to construct the correct set of symbols. If there is more disorder, or entropy, then more information is required to reconstruct the correct set of symbols.

The entropy of text is a hard thing to measure. Language is not a stationary ergodic source as information theory requires. Its probability distribution depends not only on time, but on the source, subject matter, audience, etc.

## Shannon’s N-gram Estimates

Shannon deﬁned the Ngram entropy as an estimator for the entropy of text which takes local context into account when estimating entropy by looking at N symbols at a time. An Ngram is a sequence of N symbols, NGRAM = . The symbols are taken from the set of English letters or words. If and , then the Ngram entropy estimate is given by,

This equation is the same as the conditional entropy, . Using this fact and the chain rule for conditional entropy, we get

for small values of can be calculated from standard tables of letter, digram and trigram frequencies. If spaces and punctuation are ignored we have a twenty-six letter alphabet and may be taken to be, or bits per letter. involves letter frequencies and is given by

The digram approximation gives the result

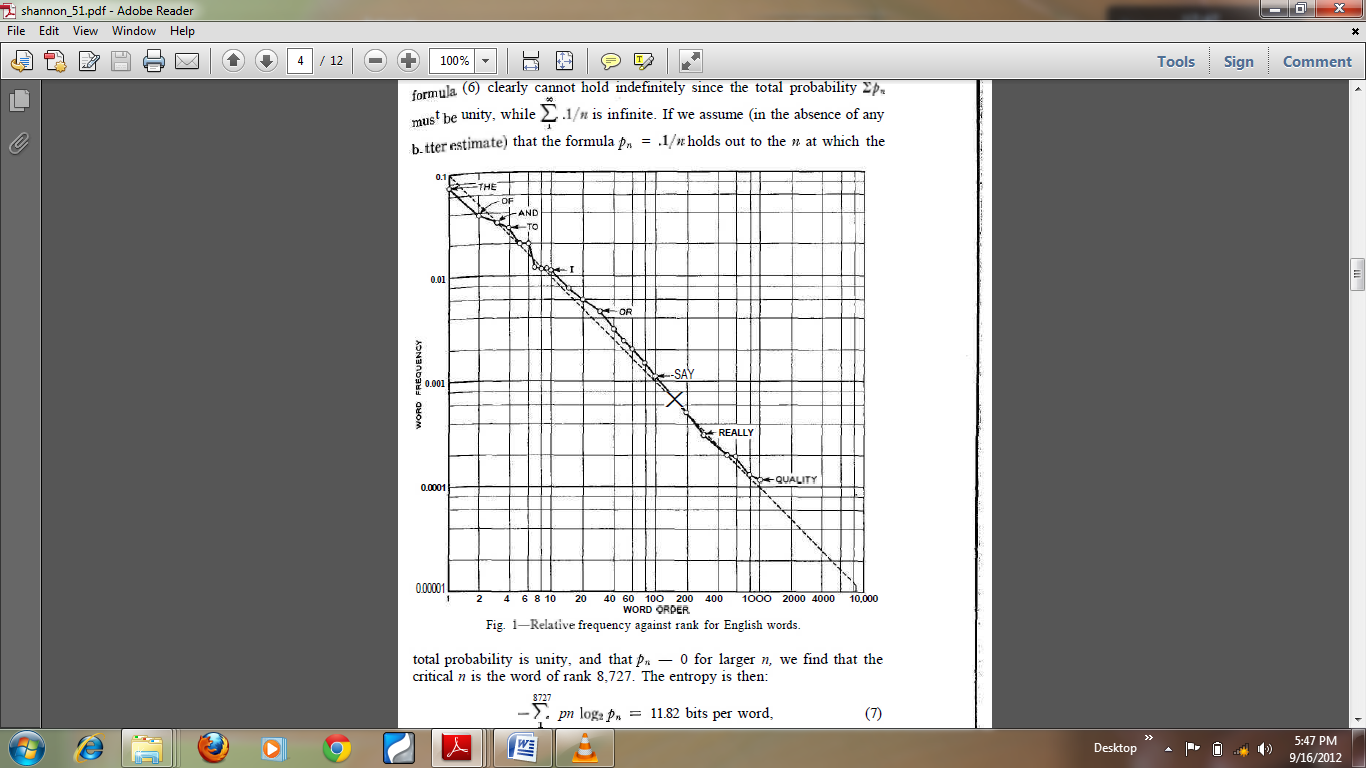
The trigram entropy is given by

In this calculation the trigram table used did not take into account trigrams bridging two words, such as WOW and OWO in TWO WORDS. To compensate partially for this omission, corrected trigram probabilities were obtained from the probabilities of the table by the following rough formula:

where is the probability of letter as the terminal letter of a word and is the probability of as an initial letter. Thus the trigrams within words (an average of 2.5 per word) are counted according to the table.

Shannon did not estimate the entropy of words using this method.

Figure 1 is a plot of the probabilities of words against frequency rank (in log).



The most frequent English word "the" has a probability and this is plotted against 1. The next most frequent word "of" has a probability of and is plotted against, etc. Using logarithmic scales both for probability and rank, the curve is approximately a straight line with a ; thus, if is the probability of the most frequent word, we have,

gives a good approximation to the word probabilities in many different languages. The formula clearly cannot hold indefinitely since the total probability must be unity, while is infinite. If we assume that the formula holds out to the at which the total probability is unity, and that for larger we find that the critical is the word of rank 8,727. The entropy is then:

Since the average word length in English is letters, this entropy comes down to 2.62 bits per letter. One might be tempted to identify this value with but actually the ordinate of the curve at will be above this value. A similar set of calculations was carried out including the space as an additional letter, giving a letter alphabet. The results of both and -letter calculations are summarized in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 26 letter | 4.70 | 4.14 | 3.56 | 3.3 | 2.62 |
| 27 letter | 4.76 | 4.03 | 3.32 | 3.1 | 2.14 |

## Shannon’s Human Estimate (Prediction of English)

Shannon conducted an interesting experiment in his 1951 paper to estimate the bounds of the entropy of text. First he assumes that the speaker of a language “possesses, implicitly, an enormous knowledge of statistics of the language. Familiarity with the words, idioms, cliches and grammar...” In the experiment, human subjects were shown 0 ≤ N ≤ 14 letters from a text unfamiliar to the subject and were asked to guess the next letter. If wrong, the subject guessed again until they got the letter right. Shannon recorded the number of guesses required for each letter. These recorded numbers, called the reduced text, carry the same amount of information as the original text. Spaces were included as an additional letter, making a alphabet. The first line is the original text; the second line contains a dash for each letter correctly guessed. In the case of incorrect guesses the correct letter is copied in the second line.

1. THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
2. -------ROO----------NOTV--------I----------SM-------OBL------
3. READING LAMP ON THE DESK SHED GLOW ON
4. REA-----------------0------------D------SHED-GLO---O--
5. POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
6. P--L--S-----------O----BU---L-S---O----------SH---------RE---C

Of a total of 129 letters, 89 or 69% were guessed correctly. The errors occur most frequently at the beginning of words and where the line of thought has more possibility of branching out. It might be thought that the second line, which we call the reduced,contains much less information than the first. Both lines contain same information in that it is possible to recover the first line from the second. To accomplish this we need an identical twin of the individual who produced the sequence. The twin will respond in the same way when faced with the same problem. The need for an identical twin in this conceptual experiment can be eliminated as follows. Good prediction does not require knowledge of more than preceding letters of text, with fairly small. There are only a finite number of possible sequences of letters. We could ask the subject to guess the next letter for each of these possible -grams. The complete list of these predictions could then be used both for obtaining the reduced text from the original and for the inverse reconstruction process.

The figure below shows a communication system constructed in which only the reduced text is transmitted from one point to the other.

Predicator

Comparison

Comparison

Original text

Original text

Reduced text

Figure : Communication System Using Reduced Text

As before, the subject knows the text up to the current point and is asked to guess the next letter and incase he is wrong, he asked to guess again until he finds the correct letter. A typical result with this experiment is shown below. The first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

1. T H E R E I S N O R E V E R S E O N A M O T O R C Y C
2. 1 1 1 5 1 1 2 1 1 2 1 1 151 1 7 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1
3. F R I E N D O F M I N E F O U N D T H I S O U T
4. 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
5. R A T H E R D R A M A T I C A L L Y T H E O T H E R
6. 4 1 1 1 1 1 1 11 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 1 (9)

Out of 102 symbols the subject guessed right on the first guess times on the second guess 8 times, on the third guess times, the fourth and fifth guesses each and only eight times required more than five guesses. Results of this order are typical of prediction by a good subject with ordinary literary English. Newspaper writing, scientific work and poetry generally lead to somewhat poorer scores.

In order to determine how predictability depends on the number of preceding letters known to the subject, a more involved experiment was carried out. One hundred samples of English text were selected at random from a book, each fifteen letters in length. The subject was required to guess the text, letter by letter, for each sample as in the preceding experiment. The following results were obtained and summarized in the table below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 100 |
| 1 | 18.2 | 29.2 | 36 | 47 | 51 | 58 | 48 | 66 | 66 | 67 | 62 | 58 | 66 | 72 | 60 | 80 |
| 2 | 10.7 | 14.8 | 20 | 18 | 13 | 19 | 17 | 15 | 13 | 10 | 9 | 14 | 9 | 6 | 18 | 7 |
| 3 | 8.6 | 10.0 | 12 | 14 | 8 | 5 | 3 | 4 | 9 | 4 | 7 | 7 | 4 | 9 | 5 |  |
| 4 | 6.7 | 8.6 | 7 | 3 | 4 | 1 | 4 | 6 | 4 | 4 | 5 | 6 | 4 | 3 | 5 | 3 |
| 5 | 6.5 | 7.1 | 1 | 1 | 3 | 4 | 3 |  | 1 | 6 | 5 | 2 | 3 |  |  | 4 |
| 6 | 5.8 | 5.5 | 4 | 5 | 2 | 3 | 2 |  |  | 1 | 4 | 2 | 3 | 4 | 1 | 2 |
| 7 | 5.6 | 4.5 | 3 | 3 | 2 | 2 | 8 |  | 1 | 1 | 1 | 4 | 1 |  | 4 | 1 |
| 8 | 5.2 | 3.6 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |  | 2 | 1 | 3 |  |
| 9 | 5.0 | 3.0 | 4 | 1 | 5 | 1 | 4 |  | 2 | 1 | 1 | 2 |  | 1 |  |  |
| 10 | 4.3 | 2.6 | 2 | 3 | 3 |  | 3 | 1 |  |  |  |  | 2 |  |  |  |
| 11 | 3.1 | 2.2 | 2 |  | 2 | 1 |  |  | 1 | 3 |  | 1 | 1 | 2 | 1 |  |
| 12 | 2.8 | 1.9 | 4 |  | 2 | 1 | 1 | 1 |  |  | 2 | 1 | 1 |  | 1 | 1 |
| 13 | 2.4 | 1.5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |  | 1 | 1 |  |  |  |
| 14 | 2.3 | 1.2 |  | 1 |  |  | 1 |  |  |  |  | 1 |  |  |  | 1 |
| 15 | 2.1 | 1.0 | 1 | 1 |  |  |  |  |  |  | 1 | 1 | 1 |  |  |  |
| 16 | 2.0 | 0.9 |  |  |  |  | 1 |  |  | 1 |  |  |  |  | 1 |  |
| 17 | 1.6 | 0.7 | 1 |  | 2 | 1 | 1 |  |  |  | 1 |  | 2 | 2 |  |  |
| 18 | 1.6 | 0.5 |  |  |  |  |  |  |  |  |  |  |  |  | 1 |  |
| 19 | 1.6 | 0.4 |  |  | 1 | 1 |  |  | 1 |  | 1 |  |  |  |  |  |
| 20 | 1.3 | 0.3 |  | 1 |  | 1 | 1 |  |  |  |  |  |  |  |  |  |
| 21 | 1.2 | 0.2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 22 | 0.8 | 0.1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 23 | 0.3 | 0.1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 24 | 0.1 | 0.0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 26 | 0.1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 26 | 0.1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 27 | 0.1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

The column corresponds to the number of preceding letters known to the subject plus one; the row is the number of the guess. The entry in column at row 5 is the number of times the subject guessed the right letter at the guess when letters were known. For example, the entry 19 in column 6, row 2 means that with five letters known the correct letter was obtained on the second guess nineteen times out of the hundred. The first two columns of this table were not obtained by the experimental procedure outlined above but were calculated directly from the known letter and digram frequencies.

One experiment was carried out with "reverse" prediction, in which the (subject guessed the letter preceding those already known. Although the task is subjectively much more difficult, the scores were only slightly poorer. Thus, with two 101 letter samples from the same source, the subject obtained the following results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. of guesses | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | >8 |
| Forward | 70 | 10 | 7 | 2 | 2 | 3 | 3 | 0 | 4 |
| Reverse | 66 | 7 | 4 | 4 | 6 | 2 | 1 | 2 | 9 |

Incidentally, the -grarn entropy for a reversed language is equal to that for the forward language as may be seen from the second form in equation. Both terms have the same value in the forward and reversed cases.

These are interesting ways to estimate entropy. Calculating entropy is all about estimating the probabilities p(x) of symbols from some source. Shannon could not adequately estimate these probabilities for English text using the mathematical tools and statistical information available at the time. So, he turned to the best processor of language available and used it to produce reduced text. Using the reduced text, he was able to estimate p(x) and entropy.

# Conclusion

It is no surprise to find out that Shannon is one of the most-cited authors in information science. We often hear Claude Shannon called the father of the Digital Age. In the beginning of his paper

Shannon acknowledges the work done before him, by such pioneers as Harry Nyquist and RVL. Hartley at Bell Labs in the 1920s. Though their influence was profound, the work of those early pioneers was limited and focussed on their own particular applications. It was Shannon’s unifying vision that revolutionized communication, and spawned a multitude of communication research that we now define as the field of Information Theory.

From this analysis it appears that, in ordinary literary English, the long range statistical effects (up to 100 letters) reduce the entropy to something of the order of one bit per letter, with a corresponding redundancy of roughly 75%. The redundancy may be still higher when structure extending over paragraphs, chapters, etc. is included. However, as the lengths involved are increased, the parameters in question become more erratic and uncertain, and they depend more critically on the type of text involved.

# References

|  |  |
| --- | --- |
| [1] | C. E. Shannon., "A mathematical theory of communication.," *Bell System Technical Journal*, 1948. |
| [2] | C. E. Shannon., "Prediction and entropy of written english," *Bell System Technical Journal*, 1951. |
| [3] | J. Fiﬁeld, "Prediction and Entropy of English Text," University of Colorado, 2006. |
| [4] | Aftab, Cheung, Kim, Thakkar, and Yedanapudi, "INFORMATION THEORY AND THE DIGITAL AGE," MIT, 2001. |
| [5] | Entropy and Redundancy In English. [Online]. <http://people.seas.harvard.edu/~jones/cscie129/papers/stanford_info_paper/entropy_of_english_9.htm> |