Error Detection and Correction - The Structure of (Natural) Languages

Eric Filiol

ESIEA - Laval
Laboratoire de cryptologie et de virologie opérationnelles
$(C + V)^O$
filiol@esiea.fr

2013 - 2014

Agenda

Agenda

Introduction : Natural Languages as Mathematical Sources

- In the Information Theory course we have presented various mathematical models of sources (DMC, stationary sources, Markov sources...).
- Natural languages or even programming languages are so complex information sources that an exact mathematical model of them is impossible.
- We however need to work practically on natural languages or on programming languages. Is it possible to build a reasonable approximation, using the concepts and tools of Information Theory ?
- Wlog, we focus on natural languages : 27-letter alphabet (26 letters + space). This extend to other languages with any alphabet.
- The core concept/tool will be that of *n-th order approximation*.

N-th Order Approximation of Natural Languages

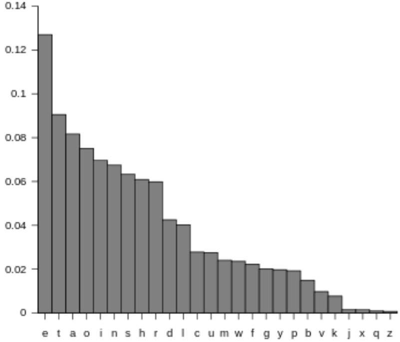
Definition

Let $\mathcal{S}$ be a Markov source of order n and $\mathcal{L}$ a (natural) language. Whenever $\mathcal{S}$ uses the distribution of n-grams of $\mathcal{L}$, we said that $\mathcal{S}$ is a n-th order approximation of $\mathcal{L}$.

- 0-th order approximation : each letter has probability $\frac{1}{27}$.
- 1-th order of approximation : letters are emitted randomly according to $\mathcal{L}$ symbols frequency
- 2-th order of approximation (bigram frequencies) or using probabilities $P(i|j) = p(i,j)/p(j)$ with respect to letters i and j.
- ...

N-th Order Approximation of Natural Languages (2)

Letter	Relative frequency in the English language
e	12.702%
t	9.056%
a	8.167%
o	7.507%
i	6.966%
n	6.749%
s	6.327%
h	6.094%
r	5.987%
d	4.253%
l	4.025%
c	2.782%
u	2.758%
m	2.406%
w	2.360%
f	2.228%
g	2.015%
y	1.974%
p	1.929%
b	1.492%
v	0.978%
k	0.772%
j	0.153%
x	0.150%
q	0.095%
z	0.074%



N-th Order Approximation of Natural Languages (3)

Guess which $\mathcal{L}$ and n have been used hereafter.

- XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD
- OCRO HLI RGWR NMIELWIS EU L NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL...
- HE AREAT BEIS HEDE THAT WISHBOUT SEED DAY OFTE AND HE IS FOR...
- IENEC FES VIMONILLITUM M ST ER PEM ENIM PTAUL
- MAITAIS DU VEILLECALCAMAIT DE LIEU DIT
- DU PARUT SE NE VIENNER PERDENT LA TET

N-the order Word Approximation of Natural Languages

- Shannon proposed to model $\mathcal{L}$ as a source whose alphabet is made up of basic words of $\mathcal{L}$ (using then word frequencies).
- 1st order word approximation of English : `REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME...`
- 2nd order word approximation of English : `THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT...`
- We can then model natural languages with a succession of Markov sources which therefore are ergodic and with a unique steady state distribution, as described in the information theory course (part IV).
- We can then use and apply all relevant tools : entropy, word entropy and explore new concepts like language redundancy.

Agenda

Agenda

The Entropy of English : a First Approximation

- We are now going to give some estimates and interpretations of the entropy of English H_E.
- By the Shannon-McMillan Theorem (on ergodic sources), we can interpret H_E by using the following formula

$$2^{nH_E} \approx t(n) \text{ for } n \text{ large} \qquad (1)$$

  where $T(n)$ denotes the number of typical (e.g. meaningful) sequences of length n of English texts.
- The critical issue is to know H_E first to compute then $T(n)$.
- First approximation : there are 27^n possible of sequences of n symbols. Then we have

$$H_E \leq \log_2(27) = 4.76 \text{ bits per symbol}$$

The Entropy of English : Using n-th Order Approximation

- A better estimate of H_E is obtained from the 1st Order Approximation and using the different probabilities that a letter occurs

$$P[< space >] = 0.18 \quad P[E] = 0.13 \ldots\ldots$$

- We use then $H(X,Y) \leq H(X) + H(Y)$ to have a better bound since

$$H_E \leq H_E^1 = -\sum_i p_i \log_2(p_i)$$

where p_i is the probability of occurence of the i-th symbol.

- We can extend to any higher order of approximation (hereafter based on the frequency of bigrams with non-zero probability)

$$H_E \leq H_E^2 = -\frac{1}{2}\sum_i \sum_j p(i,j) \log_2(p(i,j))$$

where $p(i,j)$ is the probability of occurence of bigram (i,j).

N-grams Entropies

	26-letter Alphabet	27-letter Alphabet
H_E^0	4.70	4.76
H_E^1	4.14	4.03
H_E^2	3.56	3.32
H_E^3	3.30	3.10

TABLE: n-gram entropies (Shannon, 1951)

N-grams Entropies (2)

TABLE I. Block entropy estimates $\hat{h}_n$ in bits per character for written English, as estimated from a concatenation of several long texts of altogether $\approx 7 \times 10^7$ characters. See the text for details.

	7-bit ASCII			27 characters		
n	Eq. (5)	Eq. (8)	$N \to \infty$	Eq. (5)	Eq. (8)	$N \to \infty$
1	4.503	4.503	4.503	4.075	4.075	4.075
2	3.537	3.537	3.537	3.316	3.316	3.316
3	2.883	2.884	2.884	2.734	2.734	2.734
4	2.364	2.367	2.369	2.256	2.257	2.257
5	2.026	2.037	2.043	1.944	1.947	1.949
6	1.815	1.842	1.860	1.762	1.773	1.781

FIGURE: Results by (Schürrman & Grassberger, 1996)

The Entropy of English : Using Conditional Entropy

- An alternative approach by (Shannon, 1951) was based on estimating the conditional entropies $H(X_n|X_1, X_2, \ldots, X_{n-1}$.
- Based on the natural assumption that an intelligent human being can operate as a predictor about the n-th letter when knowing the $n-1$ previous ones.

n	Lower bound	Upper bound	n	Lower bound	Upper bound
1	3.19	4.03	9	1.0	1.9
2	2.50	3.42	10	1.0	2.1
3	2.10	3.00	11	1.3	2.2
4	1.70	2.60	12	1.3	2.3
5	1.70	2.70	13	1.2	2.1
6	1.30	2.20	14	0.9	1.7
7	1.80	2.80	15	1.2	2.1
8	1.00	1.80	100	0.6	1.3

- Similar results obtained by (Burton & Licklider, 1955).
- All the previous estimations, they strongly depends on the type of natural languages texts you use (e.g *Gadsby*, E. V. Wright's 250-page noveal never uses the letter 'e').

Agenda

Introduction

- Shannon (1951) suggested to base the estimate of English entropy on word frequency rather than letter/symbol frequency.
- Let us consider a (natural) language as a word collection $\{w_1, w_2, \ldots, w_N\}$ each occuring with probability $p(w_i)$. The word entropy is then given by

$$H_W = -\sum_{i=1}^{N} p(w_i) \log_2(p(w_i)).$$

- Shannon suggested (from Formula (1)) that the symbol entropy H_E could be approximated by

$$H_E = \frac{H_W}{l(w)}$$

where $l(w)$ is the average length of a language word.

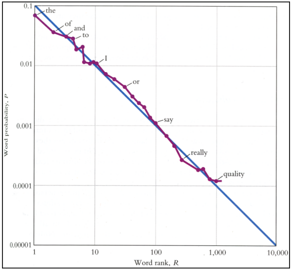
Introduction (2)

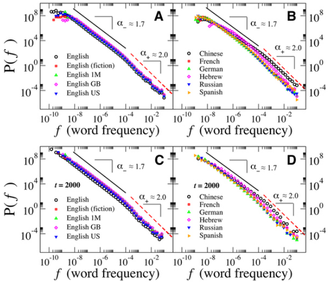
- Unfortunately, this approximation is far from being accurate since
  - Words in a language are not independent (the previous formula relates more to 1st order approximation).
  - Letters in a word are not independent.
- It is possible to have a better approximation by considering a law proposed by a linguist G.K. Zipf (1935).
- This law holds strikingly well for many various languages.
- Zipf law has been applied to many other various research area (biology, economics, earth science...).

Zipf Law

Zipf Law

The probability of occurrence of words in a language or other items starts high and tapers off. Thus, a few occur very often while many others occur rarely. Formally, if p_n is the frequency of the n-th ranked word/item (relatively to the decreasing order of their probability), then $p_n \approx \frac{A}{n}$, where A is constant that depends on the language/context in question.





Shannon's Use of Zipf Law

- Shannon used Zipf law as an approximation to the word frequencies of English, with $A = 0.1$.
- Using Zipf law with $A = 0.1$ and taking $M = 12366$, we have

$$\sum_{n=1}^{12366} p_n = 0.1 \sum_{n=1}^{12366} \frac{1}{n} = 1$$

- Then Formula (1) gives

$$H_W = 9.72 \text{ bits per word}$$

- Using this value and taking the fairly well established approximation of 4.5 letters for $l(w)$ of an English word, we obtain the estimate for $H_{4.5}$ (or equivalently the 4.5-th order approximation to English with respect to a 26-letter alphabet) :

$$H_{4.5} \approx \frac{9.72}{4.5} = 2.16 \text{ bits per letter.}$$

Discussion

- The previous approximation is an underestimate since words are actually dependent sequences of letters. Hence we have

$$H_W = \sum_{k=1}^{\infty} H(W||W| = k).P(|W| = k)$$

where W is a random word output and $|W|$ its length.

- Thus we have

$$H_W = \sum_{k=1}^{\infty} H(X_1X_2\ldots X_k).P(|W| = k) \leq \sum_{k=1}^{\infty} k.H(X).P(|W| = k),$$

where $H(X)$ is the symbol entropy H_E. The inequality follows from the basic one $H(X,Y) \leq H(X) + H(Y)$.

- This gives

$$H_W \leq H_{4.5} \sum kP(|W| = k)$$

that is

$$H_W \leq H_{4.5}.l(w)$$

Agenda

Introduction

- For (natural) languages we have related the entropy per symbol of the language with the number of meaningful messages, that is

$$T(n) \approx 2^{nH}$$

- Now by the noiseless coding theorem, a source with entropy H has a compact encoding in an alphabet Σ for typical strings of length n such that

$$l(n) \approx \frac{nH}{\log_2(|\Sigma|)} \qquad (2)$$

- We have seen that data compression in fact consists in recoding information with respect to a compact code. This implies that before compression, part of of the information is redundant and hence represents a wasted space. What is the amount of this wasted space ?

Language Redundancy

Language Redundancy

Redundancy R in information theory is the number of bits used to transmit a message minus the number of bits of actual information in the message.

- Data compression is a way to reduce or eliminate unwanted redundancy.
- Error detection/correction techniques consist of adding desired redundancy when communicating over a noisy channel of limited capacity.
- We can think of the redundancy R as a percentage. Then it is natural to write

$$l(n) \approx n(1 - \frac{R}{100})$$

- Combining with Formula (2), we have

$$R = 1 - \frac{H}{\log_2(|\Sigma|)}$$

Estimating Language Redundancy

- Estimating language redundancy precisely is very difficult. Moreover it strongly depends on the text corpus chosen.

	The Bible	William James	The Atlantic Monthly
H_0	4.086	4.121	4.152
H_1	2.397	2.654	2.824
Estimate of R	41.4 %	32.2 %	28.5 %
$l(w)$	4.06	4.556	4.653

TABLE: Data for different types of English texts (Shannon, 1951)

Estimating Language Redundancy (2)

- There is obviously significant dependence regarding the languages.

	Samoan	English	Russian
H_1 (letter)	3.37	4.114	4.612
H_{12}	2.136	2.397	2.395
Estimate of R	37.2 %	41.3 %	47.4 %
$l(w)$	3.174	4.060	5.296

TABLE: Data for identical passages of the Bible translated into different languages (Shannon, 1951)

- Samoan has a 16-letter alphabet, of which 60 % are vowels. Pre-1917 Russian used a 35-letter alphabet.

Discussion

- Shannon estimated that asymptotically the entropy of English can be reduced to something of the order of 1 bit per letter. This would correspond to a redundancy of roughly 75 %.
- This assumption has to be interpreted with care : it does not mean that it is systematically possible to recover a text in which letters are deleted with probability $\frac{1}{4}$.
- The exact nature of deletion is important (take the Arabic language as an example).
- Miller & Friedman (1957) proved that there is a critical value $p \approx 0.25$ of the deletion probability above which the recovery of the message from the mutilated text is impossible.

Discussion (2)

- While it is theoretically possible to shorten printed text to a quarter of their present length, random deletion is not a good choice.
- Big reduction can be achieved by clever and "sensible" encoding.
  - Omit space, letter R,...
  - Leave out vowels.
- The concept of redundancy is critical in cryptography (this issue will be addressed in the Cryptography Minor Course MAT5051).

Agenda

Conclusion

- According to Shannon, we have for English

$$0.5 \leq R_E \leq 0.75 \quad \text{and} \quad 1.19 \leq H_E \leq 2.38$$

- We have all concepts and tools to measure the entropy and redundancy of any language.
- This language modelling is dependent of the working corpus.
- Application : you can simulate any language and therefore mimic any sort of trafic.
- This is part of the approach in Perseus Lib upper layer (pending).
- Go now to the computer room to practice with exercices.

Agenda

Essential Bibliography

A few papers are available on the Moodle repository for this lecture.

- Burton, N. G. & Licklider, J. C. R. (1955). Long-range Constraints in the Structure of Printed English. *Amer. J. Psych.*, 68, 650–653.
- Miller, G. A. & Friedman, E. A. (1957). The reconstruction of Mutilated English Texts. *Inf. Contr.*, 1, pp. 38–55.
- Pierce, J. R. (1973). The early Days of Information Theory. *Trans. on Info. Theory*, 19, pp. 3–8.
- Schürmann, T. & Grassberger, P (1996). Entropy Estimation of Symbol Sequences. *CHAOS*, Vol. 6-3, pp. 414–427
- Shannon, C.E. (1951). Prediction and Entropy of Printed English. *Bell Syst. Tech. J.*, 30, pp. 50–64.
- Yavuz, D. (1974). Zipf's Law and Entropy. *Trans. on Info. Theory*, 20, pp. 650.
- Zipf, G. K. (1935). *The Psycho-biology of Language*. Houghton Mifflin Press.