Foundations of Computer Security

Lecture 35: Entropy of English

Dr. Bill Young
Department of Computer Sciences
University of Texas at Austin

Entropy vs. Redundancy

Intuitively, the difference between the efficiency of the encoding and the entropy is a measure of the redundancy in the encoding.

If you find an encoding with efficiency matching the entropy and *there is no redundancy*.

The standard encoding for English contains a lot of redundancy.

> *Fr xmpl, y cn prbbly gss wht ths sntnc sys, vn wth ll f th vwls mssng. Tht ndcts tht th nfrmtn cntnt cn b xtrctd frm th rmnng smbls.*

The Effect of Redundancy

The following also illustrates the redundancy of English text:

> *Aoccdrnig to rscheearch at Cmabirgde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a ttoal mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Amzanig huh?*

Spammers count on the ability of humans to decipher such text, and the inability of computers to do so to defeat anti-spam filters: *Care to order some Vi@gra or Vigara?*

Entropy of English: Zero-Order Model

Suppose we want to transmit English text (26 letters and a space). If we assume that *all characters are equally likely*, the entropy is:

$$h = -(\log 1/27) = 4.75$$

This is the *zero-order* model of English.

This gives an approximation to the entropy. But the underlying assumption is clearly false.

Entropy of English: First-Order

In written or spoken English, some symbols occur much more frequently than others.

letter	frequency	letter	frequency	letter	frequency	letter	frequency
a	0.08167	b	0.01492	c	0.02782	d	0.04253
e	0.12702	f	0.02228	g	0.02015	h	0.06094
i	0.06966	j	0.00153	k	0.00772	l	0.04025
m	0.02406	n	0.06749	o	0.07507	p	0.01929
q	0.00095	r	0.05987	s	0.06327	t	0.09056
u	0.02758	v	0.00978	w	0.02360	x	0.00150
y	0.01974	z	0.00074				

Assuming that all symbols are *independent* of one another, but follow the probabilities above, the entropy is 4.219 bits per symbol. This is the "first-order" model of English.

Entropy of English: Higher-Order

The assumption of independence (zero memory) is also incorrect. Some letters follow other letters frequently; others not at all.

The following shows the most common digrams and trigrams in English.

Digrams	**Trigrams**
EN	ENT
RE	ION
ER	AND
NT	ING
TH	IVE
ON	TIO
IN	FOR
TR	OUR
AN	THI
OR	ONE

Entropy of English: Higher-Order

We can compute tables of the likelihood of *digrams* (two-letter combinations), *trigrams*, etc. Adding digrams to the computation gives a second-order model; adding trigrams gives a third-order model; etc.

A third-order model yields 2.77 bits per symbol. The actual entropy is the "limit" of this process of taking higher and higher order models.

Estimates by Shannon based on human experiments have yielded values as low as 0.6 *to* 1.3 *bits per symbol.*

Lessons

- All natural languages contain significant redundancy.
- Computing the entropy of a natural language is difficult and requires complex models.

Next lecture: Entropy Odds and Ends