4 Density and Redundancy of a Language

SHANNON's theory provides an idea of an unbreakable cipher via the concept of perfection. Moreover it develops the concept of "unity distance" as a measure of the difference to perfection. This concept takes up the observation that the longer a ciphertext, the easier is its unique decryption.

We don't want to develop this theory in a mathematically precise way, but only give a rough impression. For a mathematiclly more ambitious approach see [11].

Unique Solution of the Shift Cipher

Let the ciphertext `FDHVDU` be the beginning of a message that was encrypted using a CAESAR cipher. We solved it by exhaustion applying all possible 26 keys in order:

Key	Plaintext	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
0	`fdhvdu`	+					
1	`ecguct`	+	+				
2	`dbftbs`	+					
3	`caesar`	+	+	+	+	+	+
4	`bzdrzq`	+					
5	`aycqyp`	+	+				
6	`zxbpxo`	+					
7	`ywaown`	?					
8	`xvznvm`	?					
9	`wuymul`	+	+				
10	`vtxltk`	+					
11	`uswksj`	+	+	?			
12	`trvjri`	+	+				
13	`squiqh`	+	+	+	+		
14	`rpthpg`	+					
15	`qosgof`	+					
16	`pnrfne`	+	+				
17	`omqemd`	+	+				
18	`nlpdlc`	+					
19	`mkockb`	+					
20	`ljnbja`	+					
21	`kimaiz`	+	+	+	?	?	
22	`jhlzhy`	+					
23	`igkygx`	+	+				
24	`hfjxfw`	+					
25	`geiwev`	+	+	+	?		

The flags in this table stand for:

- +: The assumed plaintext makes sense including the t-th letter.
- ?: The assumed plaintext could make sense including the t-th letter but with low probability.

Given the first five letters only one of the texts seems to make sense. We would call this value 5 the "unicity distance" of the cipher.

Mathematical Model

Let us start again with an n-letter alphabet Σ. The "information content" of a letter is $\log_2 n$, for we need $\lceil \log_2 n \rceil$ bits for a binary encoding of all of Σ.

Example For $n = 26$ we have $\log_2 n \approx 4.7$. Thus we need 5 bits for encoding all letters differently. One such encoding is the teleprinter code.

Now let $M \subseteq \Sigma^*$ be a language. Then $M_r = M \cap \Sigma^r$ is the set of "meaningful" texts of length r, and $\Sigma^r - M_r$ is the set of "meaningless" texts. Denote the number of the former by

$$t_r := \#M_r.$$

Then $\log_2 t_r$ is the "information content" of a text of length r or the **entropy** of M_r. This is the number of bits we need for distinguishing the elements of M_r in a binary encoding.

Remark More generally the entropy is defined for a model that assigns the elements of M_r different probabilities. Here we implicitly content ourselves with using a uniform probability distribution.

We could consider the relative frequency of meaningful texts, t_r/n^r, but instead we focus on the **relative information content**,

$$\frac{\log_2 t_r}{r \cdot \log_2 n}:$$

For an encoding of Σ^r we need $r \cdot \log_2 n$ bits, for an encoding of M_r only $\log_2 t_r$ bits. The relative information content is the factor by which we can "compress" the encoding of M_r compared with that of Σ^r. The complimentary portion

$$1 - \frac{\log_2 t_r}{r \cdot \log_2 n}$$

is "redundant".

Usually one relates these quantities to $\log_2 n$, the information content of a single letter, and defines:

Definition 2 (i) The quotient

$$\rho_r(M) := \frac{\log_2 t_r}{r}$$

is called the r**-th density**, the difference $\delta_r(M) := \log_2 n - \rho_r(M)$ is called the r**-th redundancy** of the language M.

(ii) If $\rho(M) := \lim_{r\to\infty} \rho_r(M)$ exists, it is called the **density** of M, and $\delta(M) := \log_2 n - \rho(M)$ is called the **redundancy** of M.

Remarks

1. Since $0 \le t_r \le n^r$, we have $\overline{\lim}\, \rho_r(M) \le \log_2 n$.
2. If $M_r \ne \emptyset$, then $t_r \ge 1$, hence $\rho_r(M) \ge 0$. If $M_r \ne \emptyset$ for almost all r, then $\underline{\lim}\, \rho_r(M) \ge 0$.
3. If $\rho(M)$ exists, then $t_r \approx 2^{r\rho(M)}$ for large r.

For natural languages one knows from empirical observations that $\rho_r(M)$ is (more or less) monotonically decreasing. Therefore density and redundancy exist. Furthermore $t_r \ge 2^{r\rho(M)}$. Here are some empirical values (for $n = 26$):

M	$\rho(M) \approx$	$\delta(M) \approx$
English	1.5	3.2
German	1.4	3.3

The redundancy of English is $\frac{3.2}{4.7} \approx 68\%$ (but [2] says 78%; also see [10]). One expects that an English text (written in the 26 letter alphabet) can be compressed by this factor. The redundancy of German is about $\frac{3.3}{4.7} \approx 70\%$ [10].