4 Dichte und Redundanz einer Sprache

SHANNONs Theorie bietet nicht nur durch den Begriff der Perfektheit eine Vorstellung von einer unbrechbaren Chiffre, sondern mit der „Eindeutigkeitsdistanz“ auch ein Maß für die Nähe zur Perfektheit. Dieser Begriff greift die Erfahrung auf: Je länger ein Geheimtext ist, desto leichter ist er eindeutig zu entschlüsseln. Die Theorie wird hier nicht mathematisch exakt vorgestellt; es soll nur ein Eindruck vermittelt werden. Eine mathematisch befriedigendere Theorie der Eindeutigkeitsdistanz wird in [5] entwickelt.

Eindeutige Lösung der Verschiebechiffre

Der Geheimtext `FDHVDU` sei der Beginn einer Nachricht, die mit einer CAESAR-Chiffre erzeugt wurde. Die Methode der Exhaustion bestand darin, alle 26 möglichen Schlüssel der Reihe nach anzuwenden:

Schlüssel	Klartext	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
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`	+	+	+	?		

In der Tabelle bedeutet

- +: Bis zum t-ten Buchstaben ist der Klartext noch sinnvoll.
- ?: Bis zum t-ten Buchstaben ist der Klartext mit geringer Wahrscheinlichkeit sinnvoll.

Schon ab dem vierten Buchstaben ist mit hoher Wahrscheinlichkeit nur noch einer der getesteten Klartexte möglich. Diesen Wert 4 würde man als „Eindeutigkeitsdistanz“ der Chiffre ansehen.

Mathematisches Modell

Wir starten wieder mit unserem n-buchstabigen Alphabet Σ. Der „Informationsgehalt“ eines Buchstabens ist dann ${}^2\log n$, d. h., man braucht $\lceil {}^2\log n \rceil$ Bits, um Σ binär zu codieren.

Beispiel. Für $n = 26$ ist ${}^2\log n \approx 4.7$, man braucht 5 Bits, um alle Buchstaben zu codieren. Eine solche Codierung ist z. B. der Fernschreibercode.

Sei nun $M \subseteq \Sigma^*$ eine Sprache; $M_r = M \cap \Sigma^r$ ist dann die Menge der „sinnvollen“ Texte der Länge r, $\Sigma^r - M_r$ die Menge der „sinnlosen“ Texte. Die Anzahl der ersteren wird mit

$$t_r := \# M_r$$

bezeichnet. Dann ist ${}^2\log t_r$ der „Informationsgehalt“ eines Textes der Länge r oder die **Entropie** von M_r – so viele Bits braucht man, um die Elemente von M_r in einer binären Codierung unterscheiden zu können.

> **Anmerkung.** Die Entropie wird allgemeiner für ein Modell definiert, wo die Elemente von M_r mit Wahrscheinlichkeiten gewichtet sind. Hier wurde implizit die Gleichverteilung angenommen.

Die relative Häufigkeit sinnvoller Texte, t_r/n^r, interessiert im Moment nicht so sehr wie der **relative Informationsgehalt**,

$$\frac{{}^2\log t_r}{r \cdot {}^2\log n} :$$

Für die Codierung von Σ^r braucht man $r \cdot {}^2\log n$ Bits, für die von M_r nur ${}^2\log t_r$. Der relative Informationsgehalt gibt also den Faktor an, auf den man die Codierung von M_r im Vergleich zu Σ^r komprimieren kann; der komplementäre Anteil

$$1 - \frac{{}^2\log t_r}{r \cdot {}^2\log n}$$

ist „redundant“.

Man bezieht diese Größen üblicherweise auf ${}^2\log n$ statt auf 1 und definiert:

Definition 2. (i) Der Quotient

$$\rho_r(M) := \frac{^2\log t_r}{r}$$

heißt r**-te Dichte**, die Differenz $\delta_r(M) := {}^2\log n - \rho_r(M)$ heißt r**-te Redundanz** der Sprache M.

(ii) Existiert $\rho(M) = \lim_{r\to\infty} \rho_r(M)$, so heißt $\rho(M)$ **Dichte** von M, $\delta(M) = {}^2\log n - \rho(M)$ **Redundanz** von M.

Bemerkungen

1. Es ist $0 \le t_r \le n^r$, also $\overline{\lim}\, \rho_r(M) \le {}^2\log n$.
2. Falls $M_r \neq \emptyset$, ist $t_r \ge 1$, also $\rho_r(M) \ge 0$. Falls $M_r \neq \emptyset$ für fast alle r, ist $\underline{\lim}\, \rho_r(M) \ge 0$.
3. Existiert $\rho(M)$, so ist $t_r \approx 2^{r\rho(M)}$ für große r.

Für natürliche Sprachen ist $\rho_r(M)$ – aus empirischen Beobachtungen geschlossen – im wesentlichen monoton fallend, und somit existieren Dichte und Redundanz; ferner ist stets $t_r \ge 2^{r\rho(M)}$. Empirische Werte (mit $n = 26$) sind

M	$\rho(M) \approx$	$\delta(M) \approx$
Englisch	1.5	3.2
Deutsch	1.4	3.3

Die Redundanz der deutschen Sprache entspricht $\frac{3.3}{4.7} \approx 70\%$ [4]; man kann erwarten, dass sich deutscher Text (in den 26 Buchstaben aufgeschrieben) um diese Rate komprimieren lässt. Die entsprechende Rate für Englisch ist $\frac{3.2}{4.7} \approx 68\%$ (nach [1] jedoch 78%; siehe dazu auch [4]).