Unterscheidung Code feste Länge
und solche mit variierender Länge

Feste Länge Codierungsabbildung
$c \quad A \to B^n$

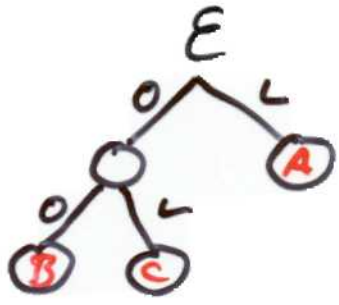
Var Länge Vorteil Häufig benutzte Zeichen können mit kurzem Wort codiert werden! Beispiel
e im Deutschen
⇒ Gesamtlänge des codierten Textes minimal
Nachteil Bei Satzcodierung ist u.U. die Trennung der Einzelwörter schwierig, vis besondere bei Störungen

Satzcodierung $c^*(\langle a_1 \quad a_n \rangle) =$
$c(a_1) \circ c(a_2) \circ \quad \circ c(a_n)$

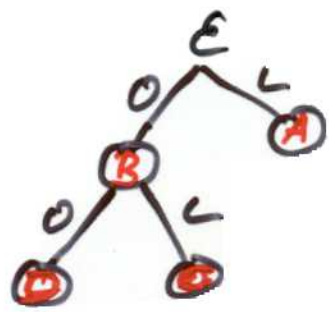
Beispiel

a) $A \to L$
$B \to 00$
$C \to 0L$

Darstellung als 'Codebaum'



b) $A \to L$
$B \to 0$
$C \to 0L$
$D \to 00$



Zu kodieren Zeichenfolge AABCDA

L L O O L O O L: bei Rekodierung möglich: A A B B A B B A, aber auch A A D A D A ..

Im Beispiel ~~Satz~~ b) ist eine (hinreichende) Voraussetzung zur Eindeutigkeit der Codierungsabbildung verletzt: Kein Codewort (CW) darf identisch mit dem Anfang eines anderen CW sein ⇒ im Codebaum sind CW nur an den Blättern ("Fano-Bedingung")

Codierung unter Berücksichtigung der Auftritts-Wk

Ziel: gegebene Nachricht einer Quelle soll durch Folge von CW codiert werden, deren Gesamtlänge Kurz.

Plausibles Vorgehen: Berücksichtigung der Häufigkeit, mit der ein bestimmtes Zeichen von der Quelle emittiert wird.
Ist über Auftritts-Wk hierzu nichts bekannt, bleibt nur, allen Zeichen gleichlange CW zuzuordnen. Beispiel: ASCII-Code mit 256 Zeichen, jedes Zeichen wird mit ld 256 = 8 bit

Allgemein gilt: Um ein Zeichen aus einem Zeichenvorrat A mit M Elementen zu codieren, werden bei Binärcodierung ld M bit benötigt; denn mit J bit lassen

sich ℓ^J Zeichen darstellen

Begriffe

a) Die Auftritts-Wk eines Zeichens a ist die Anzahl n der Auftritte des Zeichens in einer Zeichenfolge der Länge N für $N \to \infty$, also $p_a = \lim_{N \to \infty} \frac{n}{N}$

b) Der Entscheidungsgehalt eines Zeichens ist die Anzahl der Schritte, die nötig ist, ein Zeichen aus dem Zeichenvorrat zu isolieren.

c) Der Informationsgehalt eines Zeichens steht in umgekehrten Verhältnis zu seiner Auftritts-Wk. Er wird definiert als:

$I_a := ld \frac{1}{p_a} = - ld\, p_a$ in bit

d) Der von einem Zeichen der Quelle zu erwartende mittlere Informationsgehalt wird als Entropie bezeichnet. ("Überraschungswerthaltigkeit") Sie ist der Erwartungswert $H = E\{I_i\} = - \sum_{i=1}^{N} p_i\, ld\, p_i$

Je höher die Entropie der Quelle, desto größer der Informationsgehalt der von ihr ausgegebenen Nachrichten. Entropie erreicht ihr

wahrscheinlich sind und ist dann $H_0 = I_i = ld M$

Die Entropie ist niedrig, wenn die Quelle zumeist das gleiche Zeichen ausgibt. In diesem Fall reichen wenige Bits, um Nachricht zu repräsentieren

Beispiel: • Quelle mit Zeichen '1' und '0'
• Mit Auftritts-Wk p wird '1' gesendet, mit Wk $(1-p)$ wird '0' gesendet

Dann ist die Entropie

$$H = -p\, ld\, p - (1-p)\, ld\,(1-p)$$

Für $p = 0{,}5$: $H = 0{,}5 \cdot 1 + 0{,}5 \cdot 1 = 1$

$p = 0{,}25$: $H \approx 0{,}8$

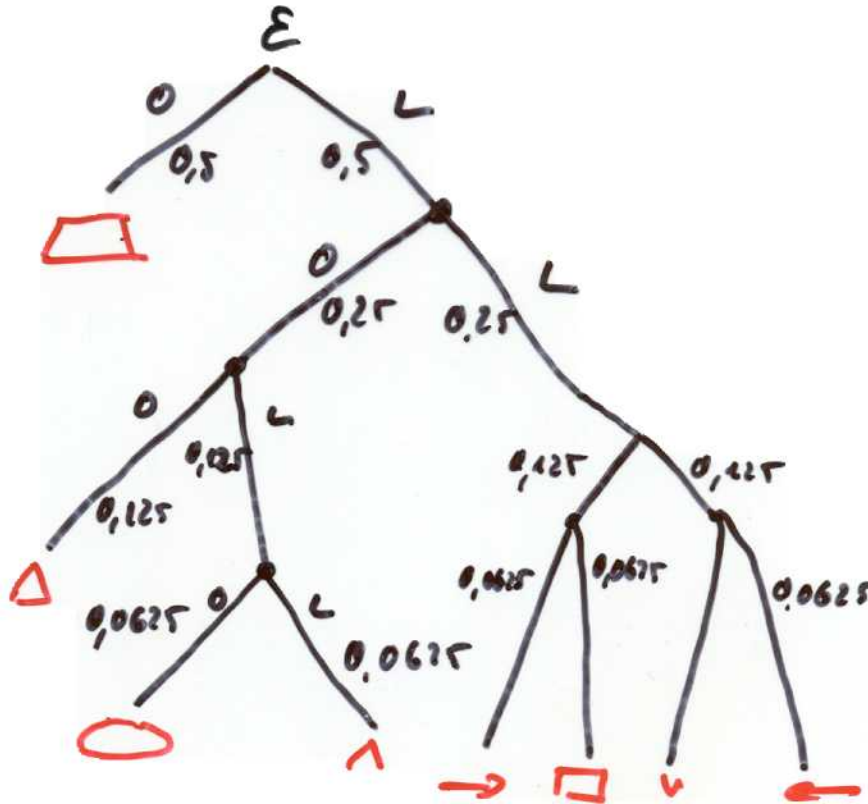
Die Länge der CW wird klein, wenn jedes Bit den maximalen Informationsgehalt transportiert, wenn also für jedes Bit die Wk für '1' bei $p = 0{,}5$ liegt $\Rightarrow$ Bit für Bit "maximale Überraschung".
Beim "Gang auf dem Codebaum" engt jedes Bit den Spielraum für mögliche Zeichen aus der Menge der CW immer weiter ein.
Bei einem (näherungsweise) optimalen Code muß daher jedes weitere Bit die noch ver-

gleich wahrscheinliche Teilmengen zerlegen

Beispiel 8 Zeichen sollen codiert bei bekannten Auftritts WK

Zeichen i	⏢	△	□	⬭	∧	∨	→	←
Pi	0,5	0,1	0,05	0,03	0,09	0,08	0,07	0,08

Eine Möglichkeit zur Codierung Konstruktion eines Codebaums von der Wurzel her aufgebaut dass sind die WK für die Restmengen der noch verbleibenden CW an jedem Knoten gleich auf die beiden Teilbäume verteilt Zuordnung der CW so dass ihre Auftritts WK möglichst nahe an der aufgeteilten "WK Masse" kommt



Huffman Codetabelle

Länge der CW ↓

Zeichen	pi	CW	pi · Li
	0,5	0	0,5
	0,1	L00	0,3
	0,05	LL0L	0,2
	0,03	L0L0	0,12
	0,09	L0LL	0,36
	0,08	LLL0	0,32
	0,07	LL00	0,28
	0,08	LLLL	0,32

$\Sigma = 2,4$ bit / Zeichen

Unbefriedigend: Abweichung der geforderten WK bei der Einsortierung in den Codebaum gegenüber den tatsächlichen WK. Dies läßt die letzte Sicherheit über die tatsächliche Wahl des Optimums vermissen.

Deshalb bessere Möglichkeit: Konstruktion des Baumes "bottom-up" ausgehend von der Beobachtung, daß die seltensten Zeichen am weitesten unten im Baum angeordnet sind. → siehe Zentralübung

Literatur

- Originalpaper von C. Shannon von 1948 → Homepage
- Heise und Querattrocchi Springer, 4 Auflage

Familie Klasse von Kompressionsalgorithmen

Beobachtung: bestimmte Zeichenfolgen häufiger als andere.

Lexikon-basierte Codierung

Literatur:

Bell et. al Modeling for Text Compression, ACM Computing Surveys, Vol. 21, No. 4, Dec. 89

Ersetzung von Zeichengruppen durch Zeiger auf "Codebuch"-Eintrag

Problem: a) Welche Zeichengruppen kommen in das Lexikon?

b) Wie wird der Text in Zeichengruppen aufgeteilt?