# Mathe III: Statistische Methoden

## Teil 2

*Garance PARIS*

*Sommersemester 2010*

# Informationstheorie

*14.–17. Juni 2010*

# Statistische Methoden

- Methoden,
  um Spracheigenschaften aus Daten zu entdecken
  - Assoziationsstärke zwischen konsekutiven Wörtern,
    Autor- oder Textarterkennung, ...
  - Methoden: Statistische Tests, ...
- Sprachmodelle,
  die versuchen, das nächste Zeichen vorherzusagen
  z. B. N-Gramm- oder Entropie-basierte Modelle
- Maschinelles Lernen:
  Klassifikation von Dokumenten, Wörtern,
  E-Mail-Nachrichten, ...

## *Informationstheorie*

- Ein Gerüst, um über den Informationsgehalt linguistischer Ereignisse nachzudenken
- Anwendungen:
  - Einen sparsamen Code finden, um Nachrichten in einer gegebenen Sprache zu senden
  - Die Fehlerwahrscheinlichkeit bei verrauschtem Kanal senken
  - Die Assoziationsstärke zwischen Wörtern beurteilen
  - Spracherkennung, OCR, Rechtschreibkorrektur, Schrifterkennung, maschinelle Übersetzung, ...
  - Daten klassifizieren
  - Sprachmodelle evaluieren

# Informationstheorie

## *Entropie (H)*

Anzahl von Bits, die benötigt werden, um den Stand eines Systems vollständig anzugeben

## *Metapher*

- Wieviele Ja/Nein-Fragen braucht man, um das Ergebnis eines Münzwurfs anzugeben?
- Wieviele braucht man bei zwei Münzen? . . . 3, 4. . . ?

# Informationstheorie

## *Entropie (H)*

Anzahl von Bits, die benötigt werden, um den Stand eines Systems vollständig anzugeben

## *Metapher*

- Wieviele Ja/Nein-Fragen braucht man, um das Ergebnis eines Münzwurfs anzugeben?
- Wieviele braucht man bei zwei Münzen? . . . 3, 4. . . ?

$$\text{Mögliche Zustände} = 2^{\text{Anzahl Fragen}}$$
$$\Leftrightarrow \log_2 \text{ Zustände} = \text{Anzahl Fragen}$$

| | |
|---|---|
| 2 Zustände | 1 Fragen |
| 4 Zustände | 2 Fragen |
| 8 Zustände | 3 Fragen |
| 16 Zustände | 4 Fragen |

# Entropie

## *Metapher, Teil 2: Verschiedene Würfel*

$\log_2 4 = 2$

$\log_2 6 = 2.585$

$\log_2 8 = 3$

$\log_2 12 = 3.585$

$\log_2 20 = 4.322$

## *Entropie bei Gleichverteilung*

In einem System mit n gleich wahrscheinlichen Zuständen gilt:

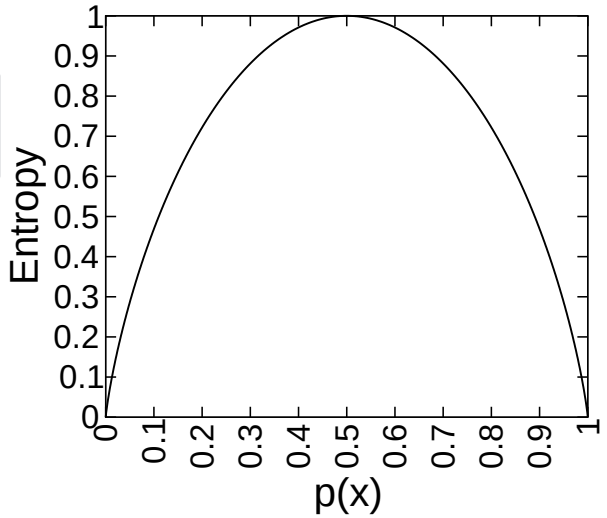
$$\begin{aligned} H &= \log_2(n) \\ &= -\log_2(1/n) \\ &= -\log_2(P(n)) \end{aligned}$$

# Entropie

## *Eine andere Sichtsweise. . .*

- Durchschnittliche Unsicherheit darüber, welches Erreignis als nächstes stattfindet
- Ein Maß dafür, wieviel Unordnung oder Zufälligkeit in einem System oder Zufallsvariable enthalten ist

## *Entropie einer gezinkten Münze*



# Entropie

## *Gesamtentropie unabhängiger Elemente*

z. B. Entropie dreier Würfel mit 4, 6, und 12 Seiten:

$$\begin{aligned} H &= \log_2(4 * 6 * 12) \\ &= \log_2(4) + \log_2(6) + \log_2(12) \\ &= 8.170 \text{ bits} \end{aligned}$$

Die Gesamtentropie ist also die *Summe* der unabhängigen Elemente im Systems

# Entropie

## *Ungleichverteilte Wahrscheinlichkeitsfunktionen*

- Die Summanden werden nach ihrer Wahrscheinlichkeit gewichtet:

$$H = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x)$$

- Wenn alle $p(x)$ in einem Korpus der Größe $N$ als $\frac{f(x)}{N}$ geschätzt werden, kann die Formel auch wie folgt umgeformt werden:

$$\frac{N \log_2 N - \sum_x x \log_2 x}{N}$$

- $0 \log 0$ wird gleich $0$ gesetzt

# Entropie

## *Bsp.: Wurf dreier Münzen*

Wie oft kommt Kopf vor,
wenn ich eine Münze dreimal werfe?

$$P(X) = \begin{cases} x = 3 & 1/8 \\ x = 2 & 3/8 \\ x = 1 & 3/8 \\ x = 0 & 1/8 \end{cases}$$

$$\begin{aligned} H(X) &= \sum_x p(x) \log_2 \frac{1}{p(x)} \\ &= \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} + \frac{3}{8} \log_2 \frac{1}{\frac{3}{8}} + \frac{3}{8} \log_2 \frac{1}{\frac{3}{8}} + \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} \\ &= 1.811 \end{aligned}$$

# Entropie

## *Weiter...*

- Untere Schranke für Anzahl von Bits, die benötigt werden, um eine Nachricht zu senden
- Eine Angabe über den Informationsgehalt einer Nachricht

# Entropie

## *Der Zustand als Nachricht*

Wie kann man das Ergebnis des Wurfs eines 8-seitigen Würfel in Bits ausdrücken?

Zum Beispiel:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 000 | 3 | 010 | 5 | 100 | 7 | 110 |
| 2 | 001 | 4 | 011 | 6 | 101 | 8 | 111 |

- Es werden 3 Bits gebraucht, um das Ergebnis als Nachricht mitzuteilen.
- Dies entspricht der Entropie: $\log_2(8) = 3$

# Entropie

## *Der Zustand als Nachricht, Teil 2*

- Bei ungleichverteilten Ereignissen kann man das Wissen um die Häufigkeit der Ereignisse mit einbeziehen, um einen besseren Code zu entwickeln
- Entropie ohne Häufigkeitsinformation: $\log_2(4) = 2$
- Der *Informationsgehalt* eines Ereignisses, der ein Teil der Entropieformel bildet, gibt die optimale Länge des zu verwendenden Codes an

$$I(p(x)) = \log_2 \frac{1}{p(x)}$$

- Bei drei Münzen:

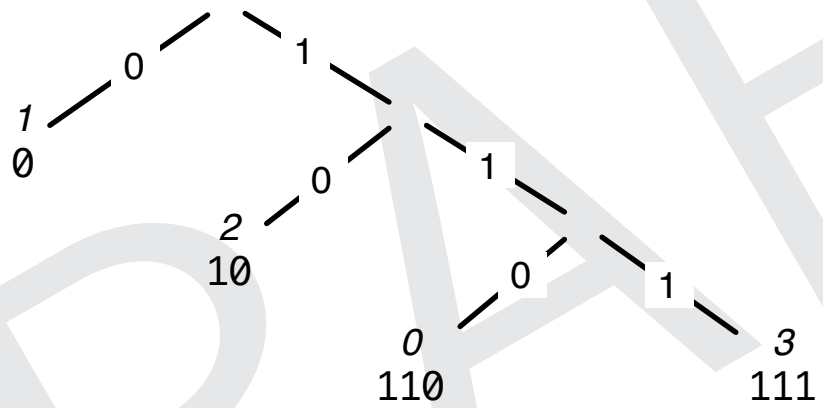
$$\text{Für } x = 0 \text{ oder } 3 \log_2 \frac{1}{1/8} = 3$$

$$\text{Für } x = 1 \text{ oder } 2 \log_2 \frac{1}{3/8} = 1.415$$

# Entropie

## *Der Zustand als Nachricht, Teil 2b*

Darauf basierend kann man einen Baum verwenden,
um einen effizienten Code für diese Sprache zu entwickeln:




Durchschnittliche Anzahl von Bits, um eine Nachricht mit diesem Code zu senden:

$$\sum_x p(x)\ \textit{code length}(x) = 3\frac{1}{8} + 2\frac{3}{8} + \frac{3}{8} + 3\frac{1}{8} = 1.875$$

# Entropie

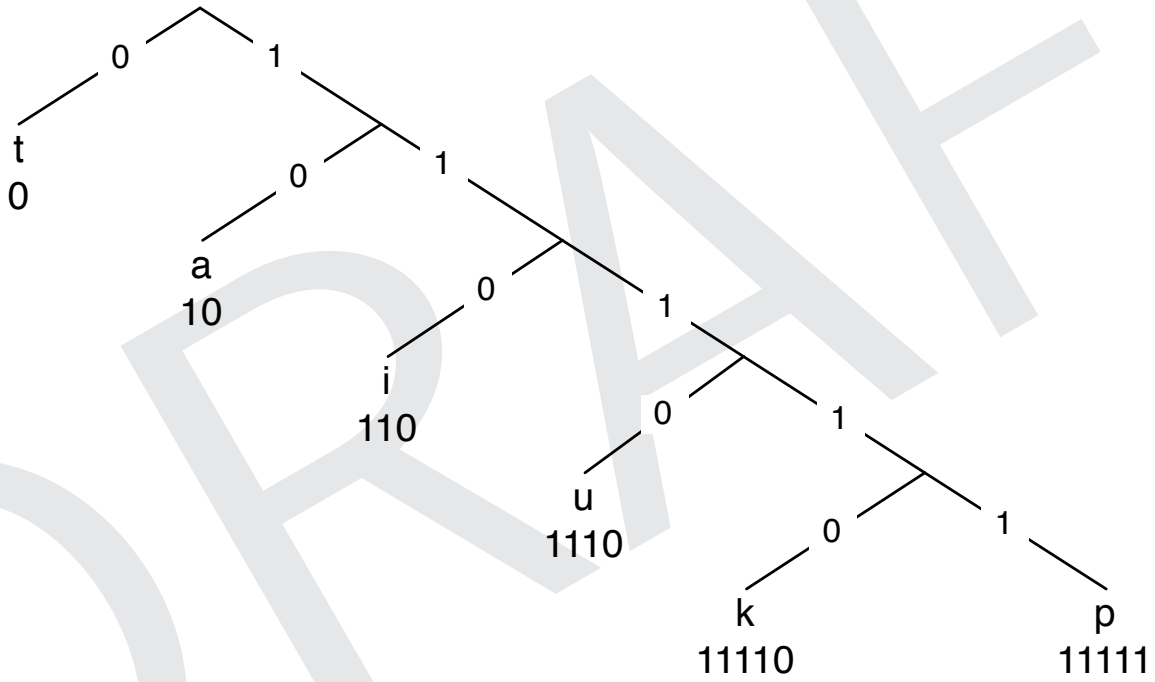
## *Beispiel: "Simplified Polynesian"*

| | |
|---|---|
| p | 1/16 |
| t | 3/8 |
| k | 1/16 |
| a | 1/4 |
| i | 1/8 |
| u | 1/8 |

$$
\begin{aligned}
H(X) &= \sum_x p(x) \log_2 \frac{1}{p(x)} \\
&= 2 * \frac{1}{16} \log_2 16 + 2 * \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 + \frac{3}{8} \log_2 \frac{8}{3} \\
&= 2.28 \text{ bits}
\end{aligned}
$$

# Entropie

## *Beispiel: "Simplified Polynesian", Teil 2*

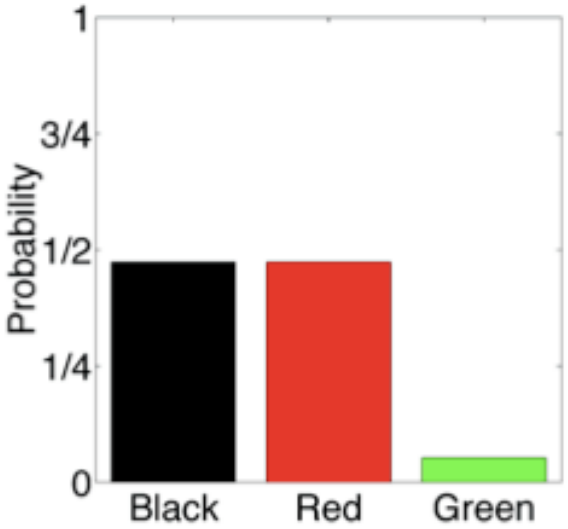


# Entropie

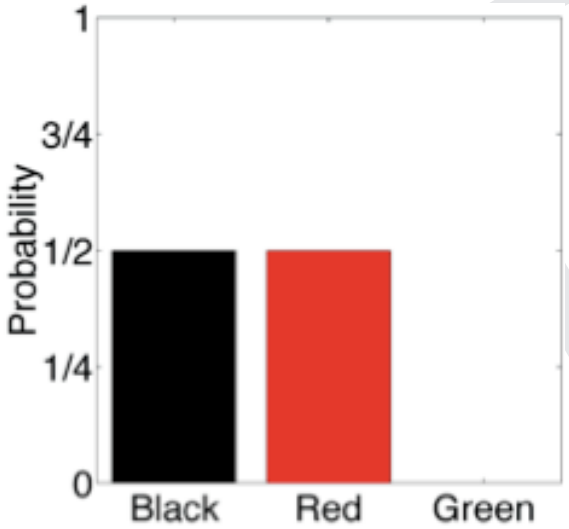
## *Einige Eigenschaften*

- Best case:
  Ein Ereignis hat die Wahrscheinlichkeit 1
  Dann ist $H = 0$
- Worst case:
  Gleichverteilung oder zufällige Sequenz von Ereignissen
- Flache, breite Verteilungen haben eine hohe Entropie
- Spitze, schmale, kompakte Verteilungen
  haben eine niedrige Entropie
- Da die Entropie unabhängiger Elemente addiert wird,
  wächst die Entropie mit der Anzahl der Dimensionen
  einer Verteilung

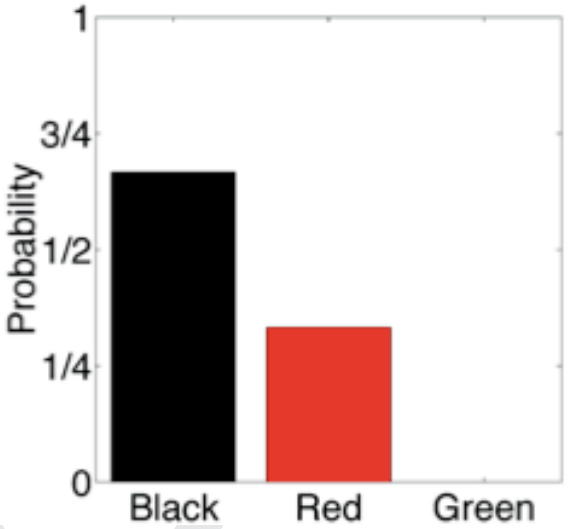
1.


1
3/4
1/2
1/4
0
Probability
Black
Red
Green

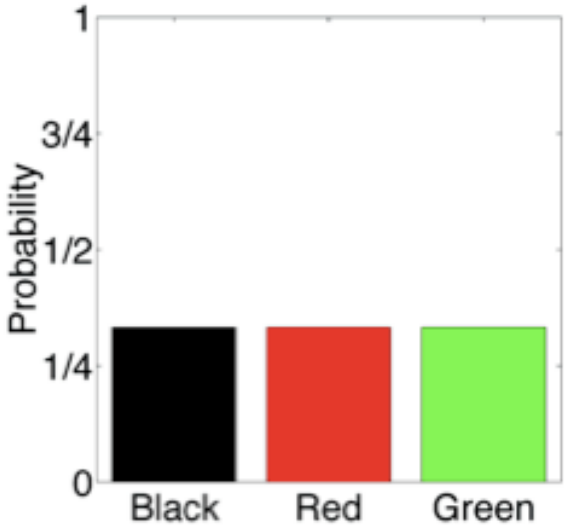
2.


1
3/4
1/2
1/4
0
Probability
Black
Red
Green

3.


1
3/4
1/2
1/4
0
Probability
Black
Red
Green

4.


1
3/4
1/2
1/4
0
Probability
Black
Red
Green

## Gemeinsame Entropie

Wieviel Information benötigt wird, um den Wert zweier Zufallsvariablen anzugeben

$$H(X, Y) = H(p(x, y)) = -\sum_x \sum_y p(x, y) \log_2 p(x, y)$$

## Bedingte Entropie

Wieviel Information benötigt wird, um Y mitzuteilen, wenn X bekannt ist

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X = x) \\ &= \sum_x p(x) \left[ -\sum_y p(y|x) \log_2 p(y|x) \right] \\ &= -\sum_x \sum_y p(x, y) \log_2 p(y|x) \end{aligned}$$

# Entropie

## *Polynesisch, Teil 2*

Jetzt finden wir aber heraus, dass alle Wörter der Sprache aus CV-Folgen bestehen

| | p | t | k | |
|---|---|---|---|---|
| a | 1/16 | 3/8 | 1/16 | 1/2 |
| i | 1/16 | 3/16 | 0 | 1/4 |
| u | 0 | 3/16 | 1/16 | 1/4 |
| | 1/8 | 3/4 | 1/8 | |

$$H(C,V) = 4 * \frac{1}{16} \log_2 16 + 6 * \frac{1}{16} \log_2 \frac{16}{3} + \frac{3}{8} \log_2 \frac{8}{3}$$
$$= 2.436$$

**pro Silbe**

# Entropie

## *Entropie Rate*

Da die Entropie von der Länge der Nachricht abhängt, ist es häufig sinnvoll zu normalisieren

$$\begin{aligned} H_{rate} &= \frac{1}{n} H(X_1, \ldots, X_n) \\ &= \frac{1}{n} H(X_{1n}) \\ &= -\frac{1}{n} \sum_x p(x_{1n}) \log_2 p(x_{1n}) \end{aligned}$$

## *Im Polynesischen...*

$2.436/2 = 1.218$ **pro Buchstabe**
(zu vergleichen mit $2.28$ ohne Wissen über Silben)

# Entropie

Entropie kann auch verwendet werden, um die Qualität eines Modells zu beurteilen

## *Relative Entropy*

- Die Relative Entropie, auch *Kullback-Leibler (KL) Divergenz* gennant, vergleicht die Entropie zweier Verteilungen
- Intuitiv: Durchschnittliche Anzahl Bits, die verschwendet werden, wenn eine Verteilung $p$ mit einem für $q$ entwickelten Code enkodiert wird
- Nicht symmetrisch!

$$D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

# Entropie

## *Kreuz-Entropie*

- Wenn wir ein Modell $m$ eines Phänomens $p$ bauen, wollen wir $D(p||m)$ niedrig halten
- Kreuzentropie:

$$H(X, m) = H(X) + D(p||m)$$
$$= \sum_x p(x) \log_2 \frac{1}{m(x)}$$

- Problem: Um die Kreuzentropie berechnen zu können, brauchen wir die Verteilung von p(x)
- Es wird angenommen, dass Sprache *ergodisch* ist, d.h. dass jede Stichprobe repräsentativ des Ganzen ist

# Entropie

## *Die Entropie von Englisch*

| | | |
|---|---|---|
| ASCII | 8 | $\log_2 256$ |
| Uniform | 4.76 | $\log_2 27$ |
| Unigram | 4.03 | |
| Bigram | 2.8 | |
| Gzip | 2.5 | |
| Trigram | 1.76 | Brown et al., 1992 |
| Human | 1.3 | Shannon *(Shannon guessing game)* |
| | 1.34 | Cover & Thomas (using gambling) |

# Mutual Information

## *Die Kettenregel*

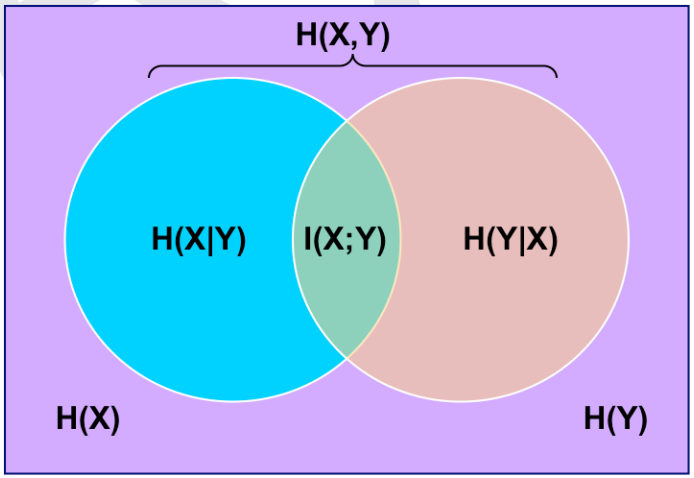
$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1})$$

## *Mutual Information*

Die Reduzierung der Unsicherheit in einer Variable, dadurch dass man über einer anderen Variable Information hat

$$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X; Y)$$



# Mutual Information

## *MI berechnen*

$$I(X;Y) =$$

$$H(X) - H(X|Y)$$

$$H(X) + H(Y) - H(X,Y)$$

$$\sum_x p(x) \log_2 \frac{1}{p(x)} + \sum_y p(y) \log_2 \frac{1}{p(y)} + \sum_{xy} p(x,y) \log_2 p(x,y)$$

$$\sum_{xy} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

## *Eigenschaften*

- Symmetrisch, immer positiv
- Maß für die Abhängigkeit zweier Verteilungen:
  $I(X;Y) = 0$ wenn X,Y unabhängig
- Wächst sowohl mit Abhängigkeit als auch mit Entropie

# Mutual Information

## *Pointwise Mutual Information*

- Maß für die Assoziationsstärke zweier Elemente
- Information,
  die in einem Ereignis $x$ über ein Ereignis $y$ enthalten ist

$$
\begin{aligned}
I(x,y) &= \log_2 \frac{p(x,y)}{p(x)p(y)} \\
&= \log_2 \frac{p(x|y)}{p(x)} \\
&= \log_2 \frac{p(y|x)}{p(y)}
\end{aligned}
$$

# Beispiel 1: Kollokationen

| Word 1 | Word 2 | C(w1) | C(w2) | C(w1 w2) | I(w1,w2) |
|---|---|---|---|---|---|
| Ayatollah | Khomeini | 42 | 20 | 20 | 18.38 |
| Agatha | Christie | 30 | 117 | 20 | 16.31 |
| cassette | recorder | 77 | 59 | 20 | 15.94 |
| unsalted | butter | 24 | 320 | 20 | 15.19 |
| over | many | 13484 | 10570 | 20 | 1.01 |
| like | people | 14093 | 14776 | 20 | 0.46 |
| time | last | 15019 | 1569 | 20 | 0.29 |

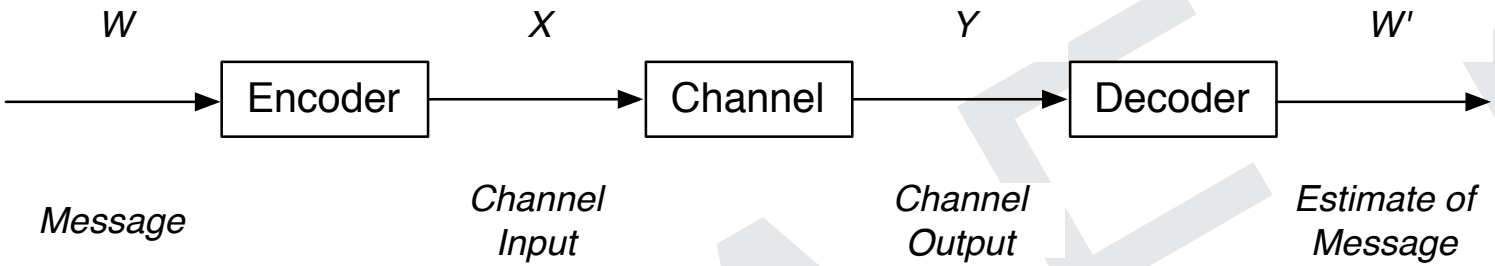
(N = 14307688)

- Pointwise MI kann verwendet werden, um mögliche Kollokationen zu bewerten:

$$I(\textit{cassette}, \textit{recorder}) = \log_2 \frac{\frac{20}{N}}{\frac{77}{N}\frac{59}{N}} = 15.94$$

- Durch das Ergebnis werden die Wortpaaren geordnet, wie bei statistischen Tests (allerdings ohne Theorie darüber, wo das Threshold zu setzen ist)

# Das Kanalmodell




- Modell der Übertragung einer Nachricht
- Eine Nachricht W wird als X encodiert
- X wird durch das Kanal übertragen
- Die Wahrscheinlichkeit der Ausgabe wird durch p(y|x) gegeben
- Y wird decodiert und als W' wiedergegeben

# Das Kanalmodell

## *Anwendungen*

Viele natürlichsprachliche Anwendungen können an Hand des Kanalmodells modelliert werden:

| Anwendung | Eingabe | Ausgabe |
|---|---|---|
| Maschinelle Übersetzung | L1 Wortfolge | L2 Wortfolge |
| OCR | Originaler Text | Text mit Fehlern |
| PoS Tagging | PoS Tags | Wörter |
| Spracherkennung | Wortfolge | Audiosignal |

# Das Kanalmodell

## *Anwendungen, Teil 2*

Es wird ein Modell gebaut, an Hand dessen man von der (bekannten) Ausgabe auf die (unbekannte) Eingabe schließen kann

$$I' = \operatorname*{argmax}_i p(i|o) = \operatorname*{argmax}_i \frac{p(i)p(o|i)}{p(o)} = \operatorname*{argmax}_i p(i)p(o|i)$$

Zwei Wahrscheinlichkeitsverteilungen müssen geschätzt werden:

- Das Sprachmodell:
  Verteilung der Zeichen in der Eingabesprache $p(i)$
- Das Kanalmodell $p(o|i)$