Elementare Wahrscheinlichkeitslehre

Vorlesung “Computerlinguistische Techniken”

Alexander Koller

13. November 2015

CL-Techniken: Ziele

- Ziel 1: Wie kann man die Struktur sprachlicher Ausdrücke berechnen?

- Ziel 2: Wie geht das effizient, auch wenn der sprachliche Ausdruck mehrdeutig ist?

- Ziel 3: Wie erkennt man unter allen möglichen Lesarten die richtige?

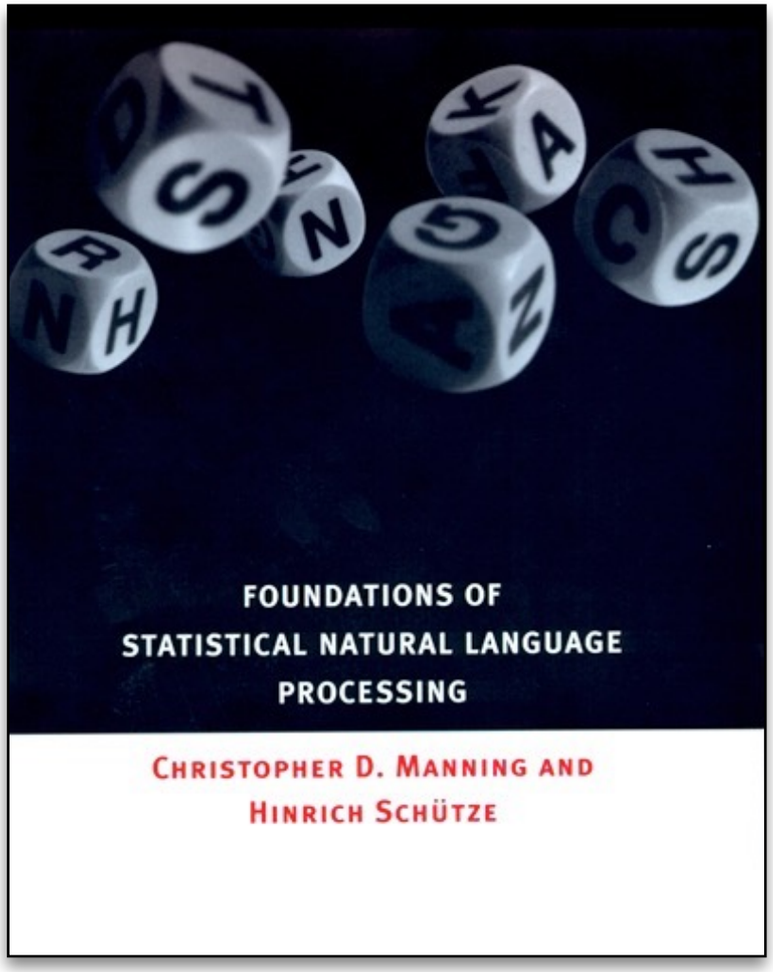
Übersicht Teil 2

- Elementare Wahrscheinlichkeitstheorie
- n-Gramm-Modelle
- Hidden Markov Models
- Probabilistische kontextfreie Grammatiken
- Einfache statistische Modelle von Semantik

Lehrbuch

- Jurafsky & Martin behandelt auch statistische Modelle.

- Etwas mehr in die Tiefe geht
  Manning & Schütze:



Heute

- Sprache als Zufallsprozess
- Elementare Wahrscheinlichkeitslehre
- Statistische Sprachmodellierung

Let’s play a game

- Wir schreiben zusammen einen Satz.
- Jeder von Ihnen diktiert mir ein Wort.
- Sie können dabei alles anschauen, was bisher an der Tafel steht.

Grundidee

- Sprache ist ein Zufallsprozess: Wörter werden Schritt für Schritt zufällig generiert.

- Viele sprachliche Phänomene sind inhärent “weich” und mit W. gut zu modellieren.
  - z.B. gradierte Grammatikalität; Selektionspräferenzen

- Umgang mit Ambiguitäten.
  - Modell: richtige Lesart = wahrscheinliche Lesart.

Wahrscheinlichkeiten

- Eine *Zufallsvariable* X, Y, … beschreibt die möglichen Ergebnisse eines Zufallsereignisses und die Wahrscheinlichkeit dieses Ergebnisses.
  - Münzwurf: Ergebnisse A = {K, Z}
    Faire Münze: P(X = K) = P(X = Z) = 0.5
    Unfaire Münze: könnte P(X = K) > P(X = Z) haben
  - Würfel: Ergebnisse A = {1, 2, 3, 4, 5, 6}
    Fairer Würfel: P(X = 1) = … = P(X = 6) = 1/6

- *Wahrscheinlichkeitsverteilung* der ZV X ist die Funktion a ↦ P(X = a).

a	P(X=a)
H	0.5
T	0.5

Ereignisse

- Die Menge A der atomaren Ergebnisse heißt der *Ergebnisraum*.

- Wir arbeiten hier mit endlichen A, d.h. *diskreten* ZV. Es gibt auch kontinuierliche ZV.

- Wahrscheinlichkeiten für komplexe Ereignisse:
  - P(X = 1 oder X = 2): W., dass X Wert 1 oder 2 annimmt.
  - P(X ≥ 4): W., dass X Wert 4, 5, oder 6 hat.
  - P(X = 1 und Y = 2): W. dass X Wert 1 und Y Wert 2 hat.

Axiome der W.theorie

- Wahrscheinlichkeiten erfüllen folgende Axiome:
  - $0 \leq P(X = a) \leq 1$ für alle Ereignisse $X = a$
  - $P(X \in A) = 1; \quad P(X \in \varnothing) = 0$
  - $P(X \in B) = P(X = a_1) + \ldots + P(X = a_n)$
    für $B = \{a_1, \ldots, a_n\} \subseteq A$

- Beispiel: Wenn X *gleichverteilt* ist mit N möglichen Ergebnissen, d.h. $P(X = a_i) = 1/N$ für alle i,
  dann ist $P(X \in B) = |B| / N$.

Rechenregeln

- Aus den Axiomen folgen einige Rechenregeln:
  - Vereinigung:
    $P(X \in B \cup C) = P(X \in B) + P(X \in C) - P(X \in B \cap C)$
  - Wenn insbesondere B und C disjunkt sind (und nur dann):
    $P(X \in B \cup C) = P(X \in B) + P(X \in C)$
  - Komplement:
    $P(X \notin B) = P(X \in A - B) = 1 - P(X \in A)$.

- Zur Vereinfachung reden wir ab jetzt nur über Ereignisse $X = a$. So gut wie alles verallgemeinert sich auf $X \in B$.

Abgekürzte Schreibweise

- Schreibweise “$P(X = a)$” ist oft unhandlich.
- Wenn die ZV klar ist, lassen wir sie weg:
  - $P(a)$ statt $P(X = a)$
  - $P(a \mid b, c)$ statt $P(X = a \mid Y = b, Z = c)$
  - $P(w_1, w_2)$ statt $P(X_1 = w_1, X_2 = w_2)$
- Es ist wichtig, sich klarzumachen, was für ZV in dieser Schreibweise gemeint sind.

Gemeinsame W.

- Wir interessieren uns oft für die W., dass zwei Ereignisse X = a und Y = b zusammen auftreten, d.h. für die *gemeinsame W.* P(X = a, Y = b).
  - z.B. X = erster Würfel, Y = zweiter Würfel
- Wenn wir gemeinsame WV kennen, können wir die einzelnen WV durch *Marginalisierung* bestimmen.

$$P(X = a) = \sum_b P(X = a, Y = b)$$

Bedingte Wahrscheinlichkeiten

- Gemeinsame W. manchmal überraschend, weil Ergebnis von X das Ergebnis von Y beeinflusst.
  - X: Ziehe eine Karte aus 52 Karten
  - Y: Ziehe danach eine zweite Karte aus dem Rest
  - P(Y ist ein As | X ist kein As) = 4/51
  - P(Y ist ein As | X ist ein As) = 3/51

- Schreibe P(Y = a | X = b) für die *bedingte W.* des Ereignisses Y = a, wenn wir wissen, dass Ereignis X = b passiert ist.

Bedingte W.

- Zusammenhang von bedingten und gemeinsamen W.:
  $P(X = a, Y = b) = P(Y = b \mid X = a) \cdot P(X = a)$
  sowie: $P(X = a, Y = b) = P(X = a \mid Y = b) \cdot P(Y = b)$

- Kann daher bedingte W. wie folgt “definieren”:

$$P(Y = b \mid X = a) = \frac{P(X = a, Y = b)}{P(X = a)}$$

$$= \frac{P(X = a, Y = b)}{\sum_{b \in B} P(X = a, Y = b)}$$

Bedingte W.

- Zusammenhang von bedingten und gemeinsamen W.:
  $P(X = a, Y = b) = P(Y = b \mid X = a) \cdot P(X = a)$
  sowie: $P(X = a, Y = b) = P(X = a \mid Y = b) \cdot P(Y = b)$

- Kann daher bedingte W. wie folgt "definieren":

$$P(Y = b \mid X = a) = \frac{P(X = a, Y = b)}{P(X = a)}$$

$$= \frac{P(X = a, Y = b)}{\sum_{b \in B} P(X = a, Y = b)}$$

Marginalisierung

Die Bayes'sche Regel

- Aus Zusammenhang von bedingter und gemeinsamer W. folgt:

$$P(X = a \mid Y = b) = \frac{P(Y = b \mid X = a) \cdot P(X = a)}{P(Y = b)}$$

- Kann damit bedingte W. "umdrehen".
  - Wir treffen jemanden mit langen Haaren (Y = L); wie w. ist es, dass es eine Frau ist (X = W)?
  - Wir nehmen an: P(Y = L | X = W) = 0.75, P(Y = L | X = M) = 0.15; außerdem P(X = W) = P(X = M) = 0.5.
  - Verwende Bayes' Regel, um P(X = W | Y = L) = 0.83 zu berechnen.

Statistische Unabhängigkeit

- Sehr wichtiger Sonderfall:
  - $P(X = a \mid Y = b) = P(X = a)$
  - äquivalent: $P(X = a, Y = b) = P(X = a) \cdot P(Y = b)$
- Ergebnis von Y beeinflusst Ergebnis von X nicht.
  Ereignisse sind *statistisch unabhängig*.
- Typische Beispiele: Münzen, Würfel
- Viele Ereignisse in Sprache sind nicht unabhängig;
  aber wir tun trotzdem so, um Modell zu vereinfachen.

Erwartungswerte

- Frequentistische Interpretation von W.:
  Wenn P(X = a) = p und wir das Experiment N-mal wiederholen, sehen wir ungefähr p · N-mal das Ergebnis “a”.

- Sei jetzt jedes Ergebnis “a” mit Belohnung R(a) verbunden. Welche Belohnung können wir nach N Runden erwarten?

$$E_P[R] = \sum_{a \in A} P(X = a) \cdot R(a)$$

Erwartungswerte

- Ursprünglich für Glücksspiele entwickelt.
  - Casino möchte Roulette-Spiel so einrichten, dass Erwartungswert für Spieler kleiner als 0 ist. (Und so ist es auch.)

- In CL wissen wir oft nicht, was die richtige Analyse ist, aber wir haben WV über Analysen. Nutze dann Erwartungswerte, um relevante Statistiken auszurechnen.

Statistische Modelle

- Wir möchten W.theorie verwenden, um ein *Modell* eines generativen Prozesses aus *Beobachtungen* zu schätzen.

- Beispiel: Wir werfen 100-mal eine Münze und beobachten dabei 61-mal Kopf. Ist die Münze fair?
  - Beobachtung: 61x K, 39x Z
  - Modell: ZV X sei *Bernoulli*-verteilt, d.h. es gibt zwei Ergebnisse, und es gibt Wert p, so dass $P(X = K) = p$ and $P(X = Z) = 1 - p$.
  - Wollen *Parameter* p des Modells aus Beobachtungen schätzen.

Häufigkeiten

- Wenn wir Daten beobachten, interessieren wir uns oft v.a. für *Häufigkeiten:*
  - *absolute Häufigkeit* C(a) des Ergebnisses “a”: Wie oft haben wir insgesamt “a” beobachtet?
  - *relative Häufigkeit* f(a): Anteil der Beobachtungen, in denen wir “a” gesehen haben, d.h. f(a) = C(a) / N, wobei N die Gesamtanzahl der relevanten Beobachtungen ist.

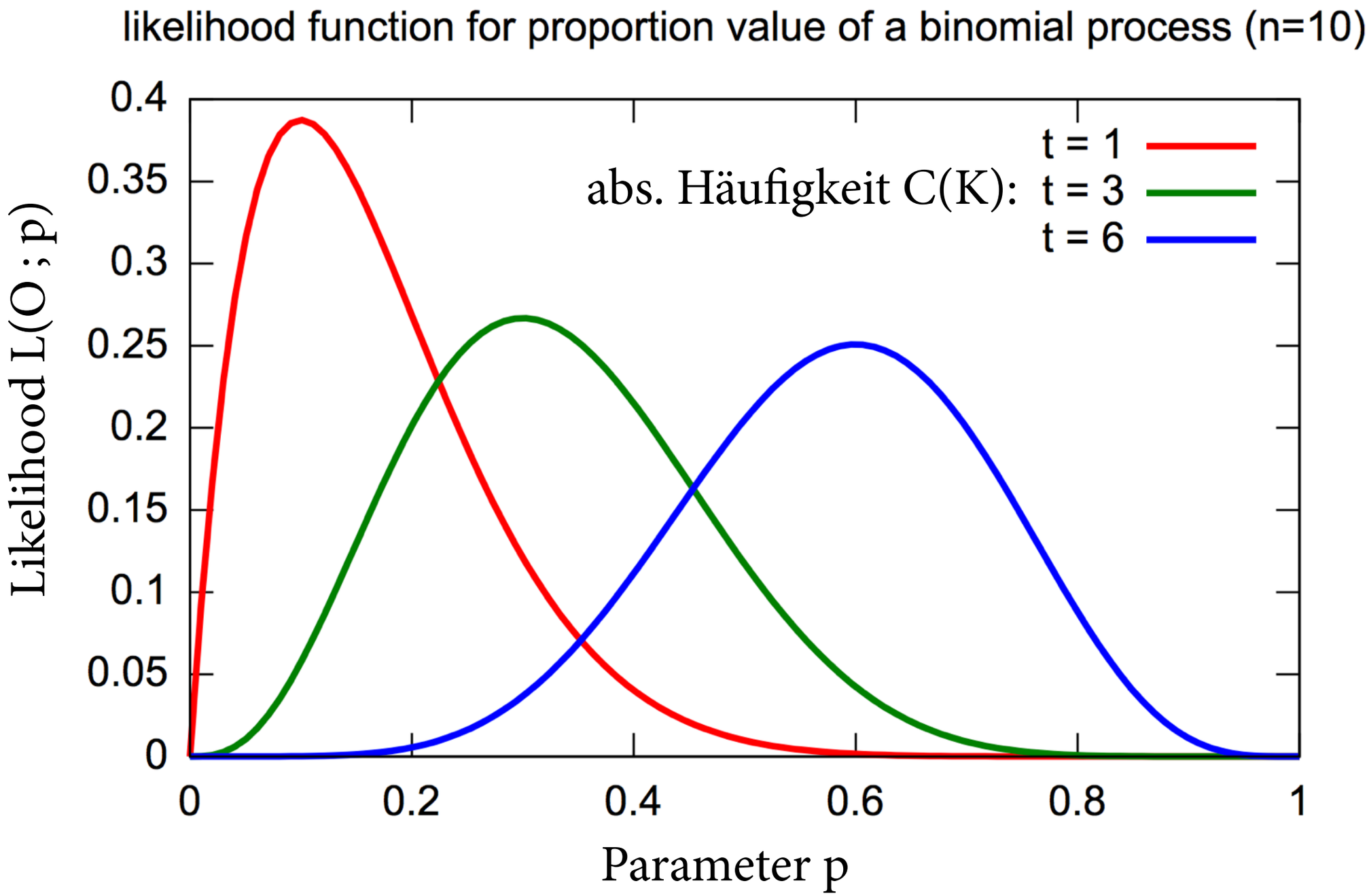
Maximum-Likelihood-Schätzung

- Kann Parameter auf viele Weisen aus Häufigkeit schätzen. Einfachster Ansatz: *Maximum-Likelihood-Schätzung*, MLE.

- *Likelihood* L(O ; p) ist die W., dass Modell Beobachtungen O generiert, wenn Parameter den Wert p hat.

- Kann man folgendermaßen ausrechnen:
  - setze P(X = a) = f(a)
  - bei Münzwurf: p = f(K)

Likelihoods



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

X_1

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

X_1

↓

Kupfer

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

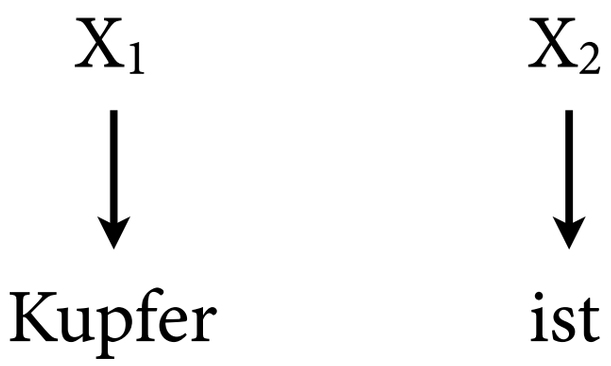
X_1 X_2

Kupfer

Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

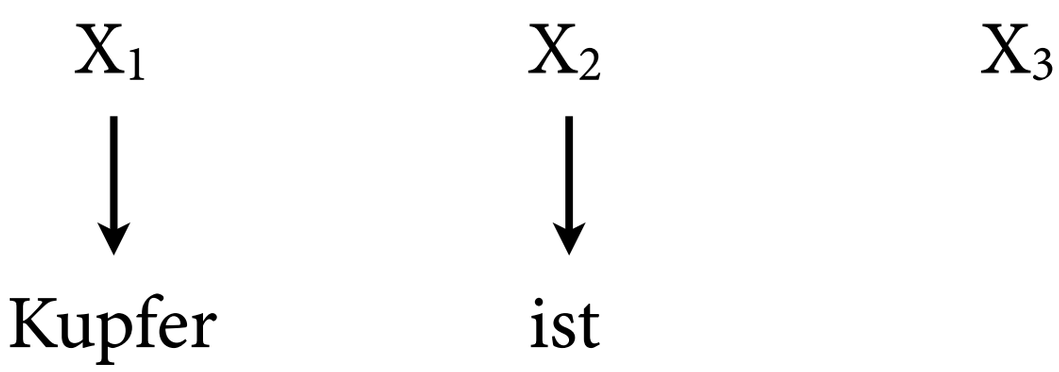
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

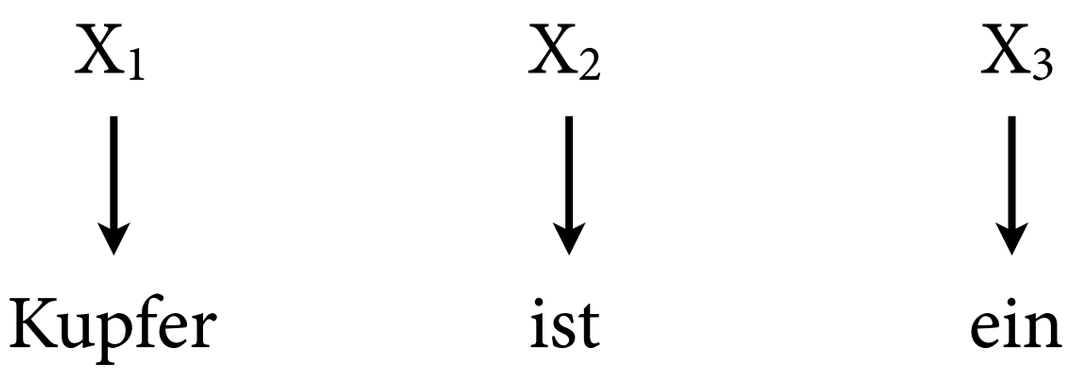
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

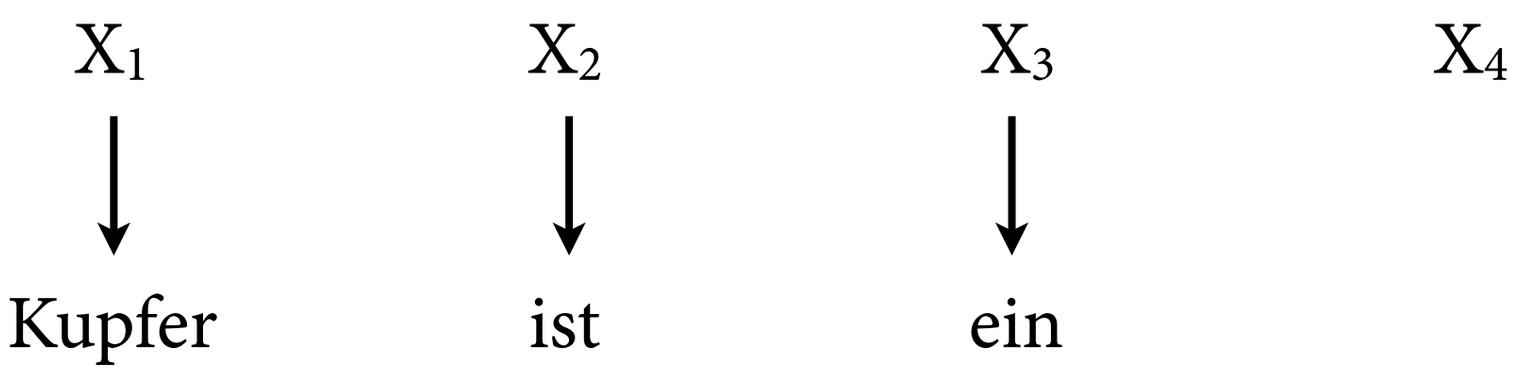
- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})



Zurück zu Sprache

- Ausgangsfrage: Nächstes Wort vorhersagen.

- Sprache als Zufallsprozess:
  - Für jede Position t im Text: ZV X_t
  - Wort w_t an Position t wird zufällig aus X_t erzeugt (abhängig von Wörtern w_1, ..., w_{t-1})

Sprache als Zufallsprozess

- Zufallsprozess wird durch W.verteilungen für die einzelnen Positionen definiert:
  - erstes Wort: $P(X_1 = w_1)$
  - zweites Wort: $P(X_2 = w_2 \mid X_1 = w_1)$
  - t-tes Wort: $P(X_t = w_t \mid X_1 = w_1, ..., X_{t-1} = w_{t-1})$

- W. des ganzen Satzes:
  $P(w_1 ... w_n) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_1, w_2)$
  $\cdot ... \cdot P(w_n \mid w_1, ..., w_{n-1})$

W. schätzen

- Zentrales Problem der statistischen CL:
  Was sind die $P(w_t \mid w_1, \ldots, w_{t-1})$?

- Schätze WV aus Beobachtung von Sprache in *Korpora*.

- Maximum-Likelihood-Schätzung:
  - $P(w_t \mid w_1, \ldots, w_{t-1}) \approx C(w_1 \ldots w_t) / C(w_1 \ldots w_{t-1})$,
    d.h. relative Häufigkeit von w_t nach Kontext $w_1 \ldots w_{t-1}$.

Einige wichtige Korpora

Korpus	Tokens	Types
Brown (NLTK)	1.1 Mio	56057
Switchboard (Englisch, gesprochen)	2.4 Mio	ca. 20.000
Penn Treebank (syntaktisch annotiert)	ca. 5 Mio	
Gigaword Corpus	1.7 Mrd	
DWDS-Korpus (deutsch)	100 Mio	
Tiger-Korpus (deutsch, syn. annotiert)	900.000	64.485

Das Zipfsche Gesetz

- Beobachtung: Die meisten Wörter kommen im Korpus selten vor.
  - in Tiger durchschnittlich 13 Tokens pro Type
  - aber: ca. 55% aller Wörter kommen nur einmal vor, ca. 70% höchstens zweimal

- Zipfsches Gesetz:
  - sortiere Wörter nach ihren absoluten Häufigkeiten
  - trage für jedes Wort absolute Häufigkeit in Graph ein
  - wenn beide Achsen logarithmisch sind, bekommt man eine Gerade

Zusammenfassung

- W.theorie ist unabdingbares Werkzeug in der modernen Computerlinguistik.

- Wichtige Konzepte heute:
  - Sprache als Zufallsprozess
  - Zufallsvariable, Wahrscheinlichkeitsverteilung
  - Unabhängigkeit, Bayes'sche Regel, Erwartungswerte
  - Statistische Modelle; Parameter; Schätzung

- Alle diese Konzepte werden immer wieder vorkommen. Fragen Sie mich bei Problemen früh!