Hilbergs Vermutung

Christian Wurm (Düsseldorf)

Düsseldorf

Hilbergs Vermutung

Hilbergs Vermutung

Hilbergs Vermutung

Hilbergs Vermutung ist, kurz gesagt:

- die bedingte Entropie natürlicher Sprache ist konstant,
- bzw. parametrisiert durch die Anzahl der Worte, die wir kennen:

Die Unsicherheit bez. des nächsten Wortes, gegeben die letzten k Worte, ist eine (polynomiale) Funktion von k.

Das bedeutet insbesondere soviel wie: die neue Information, die wir mit jedem neuen Wort bekommen, ist (mehr oder weniger) konstant – noch plakativer:

Hilbergs Vermutung

Hilberg, plakativ

Es gibt eine Konstante der menschlichen sprachlichen Informationsverarbeitung.

Was bedeutet das? Das bedeutet wir auf eine Art und Weise kommunizieren, dass der Stoff (die Info) immer auf die richtige Dichte gestreckt wird:

- Ist das Zeug zu dünn, langweilen wir uns.
- Ist das Zeug zu stark, können wir es nicht ohne weiteres verstehen!

Hilbergs Vermutung

Stochastische Prozesse

Stochastische Prozesse

Sei X eine Zufallsvariable

$$X : \Omega \to M$$

und P eine zugehörige Wahrscheinlichkeitsverteilung

$$P : \Omega \to [0, 1]$$

so dass wir

$$P(X = x) := P(X^{-1}(x)) \tag{1}$$

definieren können.

Stochastische Prozesse

Ein stochastischer Prozess ist eine **Familie** von solche Variablen

$$X_t : t \in T$$

Hier ist T die **Indexmenge**; T ist wie *time* und deswegen hier **linear**. Wir sagen

- der Prozess ist diskret, falls $T = \mathbb{N}$ oder $T = \mathbb{Z}$ etc.
- der Prozess ist stetig, falls $T = \mathbb{R}$ oder $T = \mathbb{R}_+$ etc.

In der Computerlinguistik interessieren wir uns normalerweise für diskrete Prozesse; stetige Prozesse sind eher interessant in der Physik.

Wichtig ist aber: unsere Definitionen gelten in beiden Fällen!

Stochastische Prozesse: unabhängig

Wichtig:

- die Variablen $X_t, X_{t'}$, $t, t' \in T$ sind nicht unbedingt unabhängig voneinander!

Unabhängig

Wir sagen der Prozess ist unabhängig indentisch verteilt, falls gilt: f.a. $t, t' \in T$, $t \neq t'$, $a_t, a_{t'} \in M$,

$$P(X_t = a_t) = P(X_t = a_t | X_{t'} = a_{t'}) \tag{2}$$

Sprich: alle Ergebnisse sind unabhängig, wir haben ein unabhängig iteriertes Experiment (Würfel, Münze)

Stochastische Prozesse: unabhängig

Solche Prozesse erzeugen die bekannte

- Binomialverteilung (falls Bernoulli) oder Multinomialverteilung im diskreten Fall,
- die Normalverteilung im stetigen Fall

Solche Prozesse interssieren uns hier **nicht**!

Stochastische Prozesse: Markov Eigenschaft

Manche dieser nicht unabhängig identischen Prozesse haben die **Markov Eigenschaft**. Im diskreten Fall sieht das so aus:

Markov, diskret

Der Prozess hat die Markov Eigenschaft, falls gilt: es gibt ein $n \in \mathbb{N}$ so dass

$$\begin{aligned} &P(X_{t+n} = a_{t+n} | X_{t+n-1} = a_{t+n-1}, ..., X_t = a_t) \\ &= P(X_{t+n} = a_{t+n} | X_{t+n-1} = a_{t+n-1}, ..., X_{t-k} = a_{t-k}) \end{aligned}$$

für alle t und $a_{t+n}, ..., a_{t-k}$.

Der Kompaktheit halber schreiben wir in solchen Fällen einfach:

$$P(X_{t+n} | X_{t+n-1}, ..., X_t) = P(X_{t+n} | X_{t+n-1}, ..., X_{t-k}) \tag{3}$$

Stochastische Prozesse: Markov Eigenschaft

Im stetigen Fall gilt:

Markov, stetig

Der Prozess hat die Markov Eigenschaft, falls gilt: es gibt ein $r \in \mathbb{R}$ so dass

$$\begin{aligned} &P(X_{t+r} = a_{t+r} | \{X_s = a_s : s \in [t, t+r)\}) \\ &= P(X_{t+r} = a_{t+r} | \{X_s = a_s : s \in [t-k, t+r)\}) \end{aligned}$$

oder einfach

$$P(X_{t+r} | \{X_s : s \in [t, t+r)\}) = P(X_{t+r} | \{X_s : s \in [t-k, t+r)\}) \quad (4)$$

Stochastische Prozesse: Markov Eigenschaft

Markov Prozesse sind bekannt und interessant.

- Man kann diskrete Markov-Prozesse gut modellieren mit Markov-Ketten
- Man kann auch Prozesse, die sicher *nicht* die Markov Eigenschaft haben, mit Markov Ketten approximieren –
- z.B. natürliche Sprache!

- Als ein weiteres Beispiel: Sprache generiert von eine PCFG haben nicht die Markov Eigenschaft.

Stochastische Prozesse: Abstraktere Eigenschaften

Wir interessieren uns für noch allgemeinere Prozesse. Wichtig ist hier, den Fokus nicht zu verlieren:

- Wir wollen nicht (wie oft bei Markov-Ketten) *explizit modellieren*, selbst wenn das Modell nachweislich inadäquat ist,
- sondern wir wollen prüfen, welche abstrakten Eigenschaften empirisch beobachtbare Prozesse erfüllen!

Stochastische Prozesse: Stationär

Intuitiv ist ein Prozess stationär, wenn er invariant ist unter **Verschiebung**. Das heißt:

Stationärer Prozess

$X_t : t \in T$ ist stationär, falls gilt (f.a. $t_1, ..., t_i, n$):

$$P(X_{t_1}, ..., X_{t_i}) = P(X_{t_1+n}, ..., X_{t_i+n}) \tag{5}$$

vorausgesetzt $t_1 \gg 0$ (dem Anfang der Zeit)

Sprich: die Verteilung ist gleich an allen "Zeitabschnitten", vorausgesetzt wir sind weit genug entfernt vom Anfang!

Stochastische Prozesse: Stationär

Sprich: in stationären Prozessen bleibt die Wahrscheinlichkeit, ein Ergebnis (oder eine Abfolge von Ergebnissen) an/ab einem Zeitpunkt zu treffen, ab einem gewissen Punkt gleich.

Das bedeutet insbesondere: in stationären Prozessen kann nichts passieren, was die Zukunft **dauerhaft** verändert – wir kehren langfristig immer zur alten Verteilung zurück.

(So wie bei den Simpsons, vielleicht)

Stochastische Prozesse: Absorbtion

Wir sagen ein Prozess **absorbiert** (für a), falls gilt:

$$lim_{t\to\infty} P(X_t = a) = 1 \tag{6}$$

Z.B.: unsere Markov Ketten in der CL, mit $\ltimes, \rtimes$, sind absorbierend: falls $a \neq \rtimes$, gilt

$$lim_{t\to\infty} P(X_t = a) = 0 \tag{7}$$

denn mit t (Wortlänge) konvergiert die Wahrscheinlichkeit von nicht-$\rtimes$ gegen 0!

Absorbtion, Stationarität, Markov

Theorem

Jeder Markov Prozess ist entweder absorbierend oder stationär.

Es ist klar dass er nicht beides sein kann: da

- $lim_{t\to\infty} P(X_t = a) = 0$
- aber die 0 nie erreicht wird,

müssen sich auch die anderen Wahrscheinlichkeiten stets verschieben (vergrößern).

Ergodische Prozesse

Ergodische Prozesse

Ein Prozess ist **ergodisch**, falls wir die Wahrscheinlichkeitsverteilung des ganzen Prozesses aus einem

- beliebig gewählten aber
- aber ausreichend großen

Blockabschnitt lesen können.

Alternativ und äquivalent:

Ergodische Prozesse

Die erwartete Häufigkeit eines Symbols über der Zeit konvergiert gegen die Wahrscheinlichkeitsverteilung der einzelnen Variablen X_n.

Ergodische Prozesse

Ergodische Prozesse

Die erwartete Häufigkeit eines Symbols über der Zeit konvergiert gegen die Wahrscheinlichkeitsverteilung der einzelnen Variablen X_n.

Etwas formaler heisst das: für beliebige X_t gilt

$$P(X_t = a) = lim_{N\to\infty} \frac{\sum_{i=k}^{N} P(X_i = a)}{N} \tag{8}$$

Ergodische Prozesse

Hieraus können wir ein einfaches Gegenbeispiel konstruieren: wir nehmen die Markov Kette mit

- $P(a|\ltimes) = 1$
- $P(b|a) = 1$
- $P(a|b) = 1$

Das ist weder stationär noch ergodisch:

Ergodische Prozesse

Wir nehmen die Markov Kette mit

- $P(a|\ltimes) = 1$
- $P(b|a) = 1$
- $P(a|b) = 1$

Das ist weder stationär noch ergodisch:

- Die erwartete relative Häufigkeit von a konvergiert gegen $1/2$
- Die Wahrscheinlichkeit $P(X_t = a)$ ist entweder 0 oder 1 ($\cong$ gerade/ungerade)

Das ist aber auch nicht probabilistisch!

Ergodische Prozesse

Umgekehrt: falls wir haben

- $P(a|\ltimes) = 1$
- $P(b|a) = 0.9$
- $P(a|b) = 0.9$

Diese Kette ist ergodisch;

- die anfängliche Differenz zwischen geraden und ungeraden Zahlen konvergiert gegen 0!
- dementsprechend habe ich, an einem Punkt der weit genug von 0 entfernt ist, $P(X_t = a) = 0.5$, da ich keinen Anhaltspunkt habe.

Ergodische und stationäre Prozesse

Einige intuitive Erklärungen:

- Stationär, ergodisch: Wir können von einem Ereignis $X_t = a$ keinen Rückschluss auf t ziehen (z.B. t ist ungerade)
- Ergodisch: Wir können von einem Ereignis keine Rückschlüsse auf die Vergangenheit ziehen, vorausgesetzt sie ist weit genug entfernt!

Zum Beispiel

X_1 legt die Verteilung einmal für alle Zukunft fest. Dieser Prozess ist nicht ergodisch.

Probabilistische Prozesse und natürliche Sprache

Soviel dazu. Und was hat das jetzt mit natürlicher Sprache zu tun?

Wir kommen jetzt dazu.

Was hat das mit Sprache zu tun?

Man nimmt normalerweise an, das Texte/Diskurse (nicht Sätze!) ergodische stationäre Prozesse sind.

- Jede Zufallsvariable entspricht einem Zeitpunkt
- Der Wert der sie annimmt ist das Wort/der Buchstabe das geäußert wird.

Das abstrahiert natürlich von wahnsinnig vielen Faktoren, ist aber durchaus einigermaßen plausibel.

Wortwahrscheinlichkeiten

Andras Kornai sagt:

> *Die Wahrscheinlichkeit, dass ein Wort geäußert wird, $P(w)$, ist ein unbekanntes, aber wohldefiniertes Konzept*

Das setzt, unter dieser Perspektive, voraus das Sprache als Prozess ergodisch und stationär ist! Sonst wäre die Wahrscheinlichkeit ja abhängig von

- dem Zeitpunkt, an dem sie gesagt wird oder
- irgendetwas, was vor beliebig langer Zeit gesagt wurde

Also:

Diese besprochenen Prozesse durchaus relevant, wenn wir denn annehmen, natürliche Sprache lässt sich als stochastischer Prozess auffassen!

Soviel zum Vorgeplänkel. Wir kommen nun zum Problem

Stochastische Prozesse und Information

Es ist bekannt aus Markov Ketten, dass das $n$te Ergebnis relevant ist für die Verteilung X_m, und zwar für *beliebige* n, m.

Diesen Einfluss kann man messen, und zwar mit

- der **gegenseitigen Information** von X_m, X_n, denotiert $I(X_m; X_n)$

Das gilt natürlich auch für weitere Prozesse – das schauen wir uns jetzt an!

Gegenseitige Information

Nimm an, wir haben zwei Variablen X_n, X_m. Wir definieren ihre gegenseitige Information als:

$$I(X_m; X_n) = H(X_m) + H(X_n) - H(X_m, X_n) \tag{9}$$

Also in Worten: die Summe der Entropien der Einzelverteilung minus die Entropie der Verbundverteilung.

Es gibt also folgendes: falls $P(X_m, X_n) = P(X_m)P(X_n)$ (Unabhängigkeit), dann gilt

$$H(X_m) + H(X_n) = H(X_m, X_n) \tag{10}$$

also

$$I(X_m; X_n) = 0 \tag{11}$$

Gegenseitige Information

Also:

> *Je weniger Entropie die Verbundverteilung hat ($\approx$ starke Abhängigkeit der beiden), desto mehr gegenseitige Information gibt es.*

(Das hängt zusammen mit der bedingten Entropie, definiert durch

$$H(X|Y) = \sum_{y \in Y} P(Y = y) H(X|Y = y) \tag{12}$$

Wenn wir diese Definition auflösen, bekommen wir:

$$\begin{aligned} H(X|Y) &= \sum_{x \in X, y \in Y} P(X^{-1}(x) \cap Y^{-1}(y)) log \left(\frac{P(X^{-1}(x) \cap Y^{-1}(y))}{P(Y^{-1}(y))} \right) \\ &= H(X, Y) - H(Y) \end{aligned}$$

)

Gegenseitige Information

(Im Logarithmus wird Multiplikation zu Addition, Division zu Subtraktion)

$$H(X|Y) = H(X,Y) - H(Y) \tag{13}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{14}$$

Also:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{15}$$

Also wieder eine bezaubernde Symmetrie!
)

Ergodische Prozesse

Wenn wir nun annehmen, dass X_n, X_m zu einem stationären ergodischen Prozess gehören, dann bedeutet dass, dass auch die Entropie **invariant** ist unter **Verschiebung**, sprich:

$$I(X_n, X_m) = I(X_{n+k}, X_{m+k}) \tag{16}$$

F.a. $k \in \mathbb{N}$, vorausgesetzt n, m sind groß genug.

Das bringt uns (so langsam) auf die **große Frage**.

Blöcke in Prozessen

Bislang haben wir soz. Einzelereignisse betrachtet. Wir können auch **Blöcke in Prozessen** betrachten:

$$P(X_n^{n+k} = a_n....a_{n+k}) := P(X_n = a_n, ..., X_{n+k} = a_{n+k}) \tag{17}$$

ist also einfach eine Verbundverteilung für zusammenliegende Variablen.

Wir können diese Blöcke wie einfache Variablen behandeln.

Blöcke in Prozessen

Nehmen wir zwei solche Blöcke.

$$X_n^{n+k},\ X_{n+k+1}^{n+2k}$$

Man kann sich nun fragen:

Die große Frage

Wie entwickelt sich folgende Terme:

$$lim_{k\to\infty} I(X_n^{n+k}; X_{n+k+1}^{n+2k}) \tag{18}$$

$$lim_{k\to\infty} I(X_n^{n+k}; X_{n+k+1}) \tag{19}$$

in natürlicher Sprache?

Die große Frage

Wie entwickelt sich

$$lim_{k\to\infty} I(X_n^{n+k}; X_{n+k+1}) \tag{20}$$

(NB: Wegen Invarianz ist die Größe unabhängig von n)

Intuitiv bedeutet die Frage soviel wie:

Wie vorhersehbar wird Text, je länger er wird?

Die große Frage

Es gibt hier zwei Extremfälle:

1. $lim_{k\to\infty} I(X_n^{n+k}, X_{n+k+1}) = 0$
   (Text ist rein zufällig), oder
2. $lim_{k\to\infty} I(X_n^{n+k}, X_{n+k+1}) = H(X_{n+k+1})$
   (Text wird komplett deterministisch)

Natürlich ist beides falsch: Texte sind kohärent, werden also bis zu einem gewissen Punkt vorhersagbar; aber Sprache wird niemals deterministisch.

Wir bewegen uns also zwischen den Extremen. Aber wo?

Anders formuliert

Vielleicht etwas intuitiver formuliert:

$$lim_{k\to\infty} H(X) - H(X|Y) = 0 \text{ gdw. } lim_{k\to\infty} H(X) = lim_{n\to\infty} - H(X|Y) \tag{21}$$

Die zwei Extremfälle werden dann zu

1. $lim_{k\to\infty} H(X_{n+k+1}|X_n^{n+k}) = lim_{k\to\infty} H(X_{n+k+1})$
   (Text ist rein zufällig), oder
2. $lim_{k\to\infty} H(X_{n+k+1}|X_n^{n+k}) = 0$
   (Text wird komplett deterministisch)

Die große Frage

Also nochmal:

1. $H(X_{n+k+1}|X_n^{n+k})$ sollte nicht gegen 0 konvergieren – Text wird nicht deterministisch.
2. $H(X_{n+k+1}|X_n^{n+k})$ sollte nicht gegen $H(X_{n+k+1})$ konvergieren – Test ist kohärent, Handlungen und Personen werden teilweise vorhersagbar. Anders: Text ist nicht zufällig.

Aber: wohin konvergiert der Wert dann? Das ist eine empirische Frage nach der informationellen Komplexität von Text. Aller Formalismus bis hierhin diente dem Zweck, diese Frage sinnvoll stellen zu können.

Die große Frage

Also nochmal intuitiv: je länger ein Text ist, desto "vorhersagbarer" wird er: desto eher können wir sagen was als nächstes passiert.

Vorhersagbarkeit ist invers proportional zu Information.

Die Frage nach der Information ist also einfach die Frage nach der Vorhersagbarkeit!

Erinnerung: Potenzgesetze

Ein Potenzgesetz ist eine funktionale Abhängigkeit zweier Größen (x und y) in der Form eines Monoms (Polynoms)

$$x \propto ay^b \tag{22}$$

wobei a, b die Parameter sind. Wenn also die Abhängigkeit einem Potenzgesetz folgt, bedeutet dass, es gibt a, b mit denen sie sich wie oben beschreiben lässt.

Man beachte, dass das ein symmetrischer Begriff ist!

$$x = ay^b \text{ gdw. } y = 1/x^{\frac{1}{a}} \tag{23}$$

Hilbergs Vermutung

Hilbergs Vermutung besagt nun:

Hilbergs Vermutung (intuitiv)

Zwischen $lim_{k\to\infty} I(X_n^{n+k}, X_{n+k+1})$ und k in natürlichen Sprachen besteht ein Potenzgesetz.

Sprich: die gegenseitige Information wächst als eine Potenz der Blöcke!

Hilbergs Vermutung

Hilbergs Vermutung (formal)

$$H(X_{n+k+1}|X_n^{n+k}) \propto k^{-1+\beta} \tag{24}$$

wobei β ein Parameter ist; empirisch gilt wohl $\beta \approx 0.5$ und $k \leq 100$ (also gilt die Korrelation nur in begrenztem Raum!)

Was bedeutet das?

$$H(X_{n+k+1}|X_n^{n+k}) \propto \frac{1}{k^{1-\beta}} \tag{25}$$

Hilbergs Vermutung

$$H(X_{n+k+1}|X_n^{n+k}) \propto \frac{1}{k^{1-\beta}} \tag{26}$$

Also: Die bedingte Entropie sinkt mit wachsendem k; je größer β, desto langsamer sinkt sie (zur Erinnerung: $k^{0.5} = \sqrt{k}$)

Falls $\beta = 0$, würde gelten: das $n$te Wort enthält ungefähr k mal mehr Information als das knte.

Falls $\beta = 0.5$, würde gelten: das $n$te Wort enthält ungefähr $\sqrt{k}$ mal mehr Information als das knte.

Hintergrund

Wir besprechen nun einige verwandte Dinge und Gesetze, um Hilbergs Vermutung einzuordnen.

Zipfs Gesetz

Zipfs Gesetz ist sehr einfach:

Zipfs Gesetz

Der Rang eines Wortes (Position in der Liste geordnet nach Häufigkeit) ist invers Proportional zu seiner Häufigkeit, also

$$f(w) \propto \frac{1}{R(w)} \tag{27}$$

Zipf-Mandelbrot Gesetz

Zipf-Mandelbrot ist eine Generalisierung und Parametrisierung davon:

$$ZM(k|N,q,s) = \frac{1/(k+q)^s}{H_{N,q,s}} \tag{28}$$

Hier gilt:

1. $k \in \{1, ..., N\}$ ist der Rang, also das eigentliche Argument der Funktion
2. N die Größe des Lexikons
3. q, s sind freie Parameter der Verteilung.

Weiterhin:

Zipf-Mandelbrot Gesetz

Zipf-Mandelbrot ist eine Generalisierung und Parametrisierung davon:

$$ZM(k|N,q,s) = \frac{1/(k+q)^s}{H_{N,q,s}} \tag{29}$$

Weiterhin:

$$H_{N,q,s} = \sum_{i=1}^{N} \frac{1}{(i+q)^s} \tag{30}$$

Herdans Gesetz für Worte

Dieses Gesetz beschreibt eine Abhängigkeit

$$\text{Anzahl von verschiedene Worte im Text} \sim \text{Textlänge}$$

Also im Prinzip die Relation

$$\text{Anzahl types} \sim \text{Anzahl token.}$$

- Sei $|T|$ die Textlänge (Anzahl token)
- $tp(T)$ ist die Anzahl der types.

Herdans Gesetz für Phrasen

- Sei $|T|$ die Textlänge (Anzahl token)
- $tp(T)$ ist die Anzahl der types.

Dann sieht das Gesetz wie folgt aus:

$$tp(T) = K|T|^{\beta} \tag{31}$$

wobei K, β empirisch zu bestimmende Parameter sind, mit üblicherweise

- $K \in [10, 100]$
- $\beta \in [0.4, 0.6]$

Der Lempel-Ziv Code

Zuletzt müssen wir noch kurz kodieren.

Information und Kodierung

Wir erinnern uns: je weniger Information in einem Text ist, desto kompakter kann man ihn kodieren. Informationstheoretisch ist das sogar eine **Definition**: man kann die Information in einem Text **messen** mit der kürzesten Kodierung.

Hier wird es aber delikat, denn verschiedene Kodierungen haben verschiedene Längen. Wir schauen uns eine davon an.

Wichtig ist: unser Kode darf keine Information verlieren!

Der Lempel-Ziv Code

(Meist Lempel-Ziv-Welch) Lempel-Ziv ist die Grundlage von .zip und vielen anderen (verlustfreien) Formaten!
Intuitiv funktioniert der Kode wie folgt:

- Jede Zeichenfolge (Länge 2) bekommt ein Kürzel zugewiesen.
- Bei Wiederholung wird sie durch das Kürzel ersetzt
- Das wird iteriert appliziert, so dass auch Folgen beliebiger Länge Kürzel zugewiesen bekommen (vorausgesetzt, sie bestehen aus zwei Teilfolgen, die bereits ein Kürzel haben).

Hieraus folgt: je repetitiver ein Text, desto kompakter wird er kodiert!

Der Lempel-Ziv Code

Der LZ-Kode ist wichtig für uns, weil er "dynamisch" *während* der Text verarbeitet wird entsteht.

Er gibt also ein gutes Mass nicht für die Gesamtentropie eines Textes (wie der Huffman-Code), sondern für die bedingte Entropie neuer Worte.