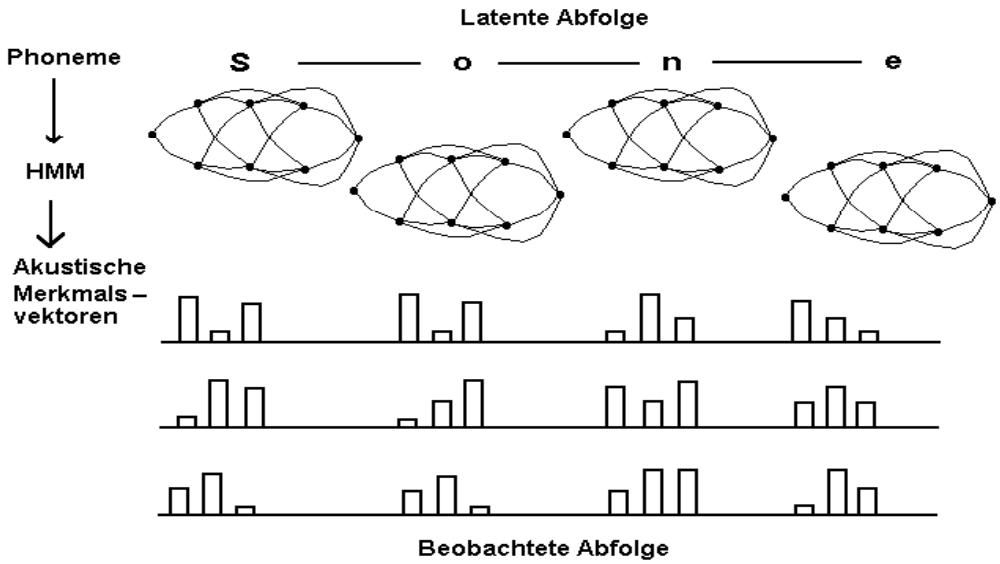
Hidden Markov Modelle

(Vorabversion begleitend zur Vorlesung Spracherkennung und integrierte Dialogsysteme am Lehrstuhl Medieninformatik am Inst. f. Informatik der LMU München, Sommer 2005)

Prof. Marcus Spies, LMU München
Diese Folien wurden vom Autor zur Einführung in das Thema 1994 verfaßt in der ...
Abteilung Spracherkennung
IBM Deutschland Informationssysteme GmbH
Wissenschaftliches Zentrum
Vangerowstr. 18
69115 Heidelberg
Geringfügige Korrekturen und Aktualisierungen wurden vorgenommen.

Warum Hidden Markov Modelle ? (1)



Erkennung: Rückschluß von beobachteter auf latente Folge (Akustik → Phonetik)

klassisches Markov-Modell:

Zustände mit Übergängen,

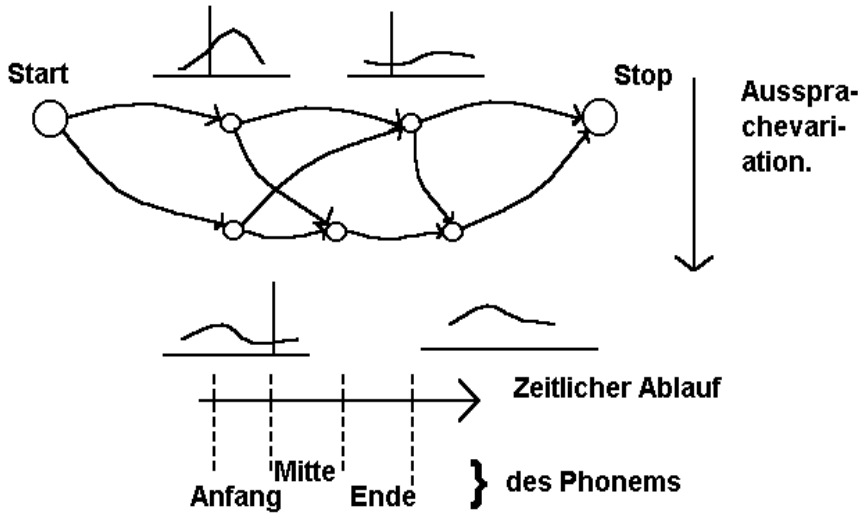
Anfangsverteilung der Zustände wird durch einen Wahrscheinlichkeitsvektor gegeben.

Übergänge werden durch Übergangsmatrix bedingter Wahrscheinlichkeiten definiert, dabei üblicherweise (!) „schwaches Gedächtnis“: nur der vorausgegangene Zustand beeinflußt die Wahrscheinlichkeit des aktuellen Zustandes

(Themen dabei: Stationarität, absorbierende Zustände, Ergodizität ...)

Warum Hidden Markov Modelle ? (2)

Schematische Modellierung von Phonemen:



(in diesem Bild werden Verweil-Transitions nicht dargestellt.)

Grundansatz

Zu modellieren ist ein Segment akustischer Daten: Unit.

Dieses Segment besteht aus mehreren akustisch zusammenhängenden Ereignissen, „kleinsten akustischen Einheiten“.

Jeder kleinsten akustischen Einheit (KAE) entspricht ein Zustand im HMM.

Kleinste akustische Einheiten unterliegen statistischen Schwankungen, die durch Emissionswahrscheinlichkeiten des jwlg. Zustands modelliert werden.

Die Verweildauer bei einer KAE wird durch eine Schleife im HMM für einen Knoten mit dem jeweiligen

Zustand modelliert – dies entspricht einem Übergang von einem Knoten zu sich selbst.

Die möglichen Übergänge von einer KAE zu einer oder mehreren weiteren KAE werden durch weitere Übergangswahrscheinlichkeiten modelliert.

Wegen der zeitlichen Eigenschaften des Sprachsignals machen i.a. rückwärtige Übergänge von KAE zu einer bereits besuchten KAE keinen Sinn. (left-to-right HMM)

Units können sein: Worte, Phoneme, Senone.

Für ASR-Anwendungen mit kleinen Vokabularen können Wortmodelle sinnvoll sein (Bsp. Kommandos, Ziffern ...).

Für ASR-Anwendungen mit großen Vokabularen müssen möglichst kleine wortunabhängige akustische Einheiten modelliert werden, daher hier eher Senon-Modellierung.

Ein Senon ist eine kleinste akustische Einheit, die sich als Blatt eines Entscheidungsbaums für Aussprachenmodellierung bilden läßt. Senone repräsentieren i.a. akustisch homogene Einheiten.

HMMs für Senone enthalten nur Schleifen und einfache Vorwärtsübergänge (d.h., Verzweigung nur bzgl. Verweilen / Verlassen je einer KAE).

Senone enthalten nur genau eine KAE, es gibt in einem entspr. HMM lediglich eine Schleife f. Modellierung der Verweilwahrscheinlichkeit).

Ein Senon kann unmittelbar gegen Merkmalsvektoren aus akustischen Daten gematcht werden.

HMMs für Units können zu größeren HMMs aggregiert werden.

Hidden Markov Modelle: Analysen

drei Fragen:

1. **Evaluation: Wahrscheinlichkeit einer beobachteten Sequenz – Ermittlung von deren Likelihood**
2. **Decoding: Ermittlung der besten latenten Sequenz anhand einer beobachteten Sequenz**
3. **Learning: Schätzung der Modellparameter anhand von Trainingssequenzen**

Hidden Markov Modelle:Grundbegriffe(1)

S **:** **Menge von Zuständen (states)**

$s \in S$ **: einzelner Zustand**

$1...,t,...T$ **: Zeitpunkte (diskret)**

O **:** **Folge beobachteter outputs (Observations)**
(z.B. O_t **: output zum Zeitpunkt** t**) ;**
o **bezeichnet einen Output.**

Hidden Markov Modelle:Grundbegriffe(2)

b_s **($s \in S$): Ausgabewahrscheinlichkeit im Zustand s. Hier speziell : $b_s(o)$: Wahrscheinlichkeit im Zustand s output o zu generieren. (Emissionswahrscheinlichkeit; hier zunächst diskret)**

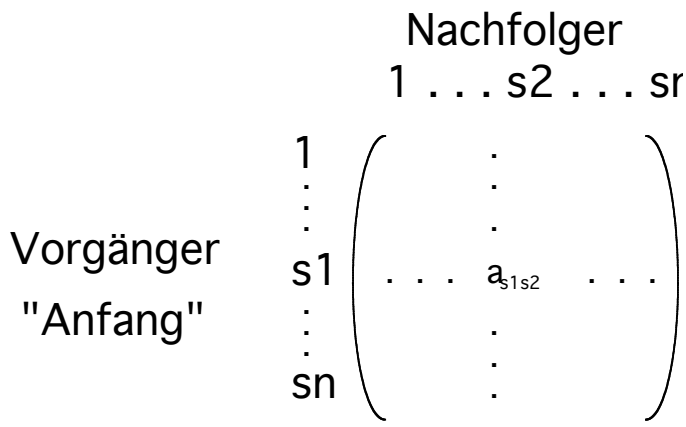
a_0 **Vektor von Anfangswahrscheinlichkeiten z.B. : a_{0s} : Anfangswahrscheinlichkeit für Zustand $s \in S$.**

(Dieser Vektor wird überflüssig, wenn man einen speziellen Startzustand annimmt, der von allen übrigen Zuständen aus mit Wahrscheinlichkeit 0 erreicht wird.)

$a_{s_1 s_2}$ **Wahrscheinlichkeit vom Zustand s_1 in Zustand s_2 zu wechseln: Übergangswahrscheinlichkeit.**

Hidden Markov Modelle:Grundbegriffe(3)

Übergangsmatrix : (stochastische Matrix) Einträge $a_{s_1 s_2}$
jeweils ≥0 und ≤ 1. Zeilensumme jeweils 1.



Hidden Markov Modelle:Grundbegriffe(4)

λ **Parameter, die die Wahrscheinlichkeiten steuern**

$P_\lambda(O)$ **Wahrscheinlichkeit, die Folge O zu beobachten, wenn λ alle Parameter steuert.**
[alternativ : $P(O|\lambda)$]

Einfaches Beispiel: $a_0 = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}$ $A = \begin{pmatrix} \lambda_{11} & . & . \\ . & . & \\ . & & \lambda_{33} \end{pmatrix}$

[Kompliziertes Beispiel :

z.B. Binomial/Multinomialverteilunng]

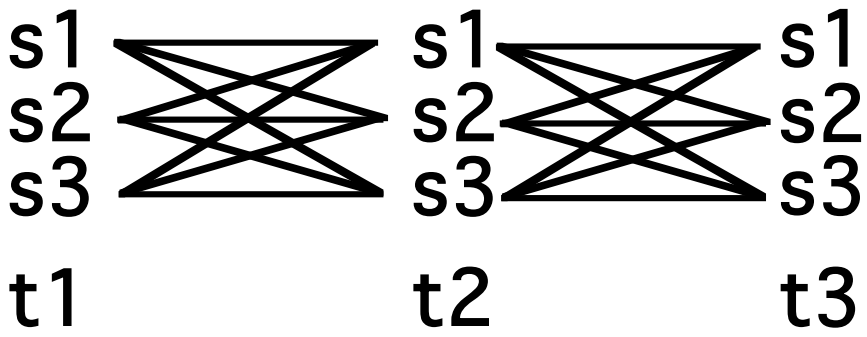
Wahrscheinlichkeit einer Beobachtungsfolge (1)

1.Frage : Wie groß ist $P_\lambda(O)$ für bestimmtes O ?
(entspricht $P(\text{"akust. Kette"}|\text{"Wortmodell"})$ im Bayesschen Klassifikator)

Antwort:

Sei S die Menge aller T-langen Pfade mit Zuständen aus S.

Bsp: $T = 3, |S| = 3.$ Dann allgemein $|\boldsymbol{S}| = |S|^T$



Unterscheidung diskrete – kontinuierliche HMM

je nach Emissionswahrscheinlichkeiten

folgende Darstellungen beziehen sich zunächst auf diskrete HMM; O_t bezeichnet das im Zeitpunkt t beobachtete (observed) Symbol

im kontinuierlichen Fall bezeichnet O_t den in t beobachteten Merkmalsvektor

Wahrscheinlichkeit einer Beobachtungsfolge (2)

$$P_\lambda(O) = \sum_{s \in S} P_\lambda(O,s) = \sum_{s \in S} P_\lambda(O|s) P_\lambda(s)$$

***s* : ein Pfad über T Zeitpunkte**

intuitiv:

Summe über alle möglichen Pfade:

(Wahrscheinlichkeit des Pfades und der beobachteten Folge)

(bedingte Wahrscheinlichkeit der Folge gegeben Pfad * Wahrscheinlichkeit des Pfades)

Wahrscheinlichkeit einer Beobachtungsfolge (3)

- $P_\lambda(O|s) = \prod_{t=1}^{T} b_{s_t}(O_t)$ **Produkt über alle Zeitpunkte :**

  Ausgabewahrscheinlichkeit beim Pfadelement s_t des Symbols O_t.

- $P_\lambda(s) = a_{0s_1} \cdot \prod_{t=1}^{T} a_{s_{t-1}s_t}$ **Produkt :**

  Anfangswert des ersten Zustandes * alle Übergangswerte der

nachfolgend aufeinanderfolgenden Zustände im Pfad *S*.

Wahrscheinlichkeit einer Beobachtungsfolge (4)

Ergebnis:

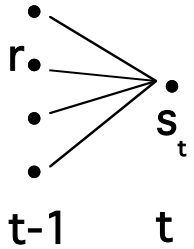
$$P_\lambda(O) = \sum_{s \in S} a_{0 s_1} \cdot b_{s_1}(O_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} b_{s_t}(O_t)$$

praktisch unbrauchbar (Komplexität in der Größenordnung $O(|S|^T)$

Vereinfachung : Forward/Backward Algorithmus.

"Forward-Variable": **Wahrscheinlichkeit der beobachteten Folge** O_1 **bis** O_t **und Zustand** s_t **zum Zeitpunkt** t**.**

Wahrscheinlichkeit einer Beobachtungsfolge (5)



$$\alpha_t(s) = P_\lambda(O_1 \ldots O_t, s_t = s)$$

Diese Variable ist rekursiv zu berechnen.

$$\alpha_t(s) = \left(\sum_{r \in S} \alpha_{t-1}(r) a_{rs} \right) b_s(O_t)$$

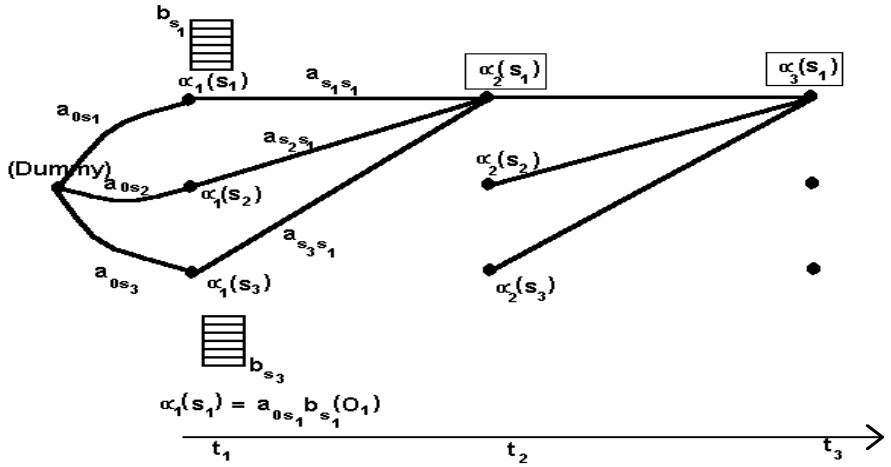
Rekursion

$$\alpha_1(s_i) = a_{0 s_i} b_{s_i}(O_1)$$

Wahrscheinlichkeit, zum ersten Zeitpunkt Zustand s_i zu haben und Symbol O_1 auszugeben.

Komplexität in der Größenordnung $O(|S|^2 T)$

Wahrscheinlichkeit einer Beobachtungsfolge (6)



Wahrscheinlichkeit einer Beobachtungsfolge (7)

Folgerung: $\sum_{s \in S} \alpha_T(s) = \sum_{s \in S} P_\lambda(O_1,...,O_T, s_T = s)$

alle Zustände

$= P_\lambda(O_1,...,O_T) = P_\lambda(O)$

Gesamtwahrscheinlichkeit beobachtete Folge

= Summe der Forward-Variablen zum letzten Beobachtungszeitpunkt.

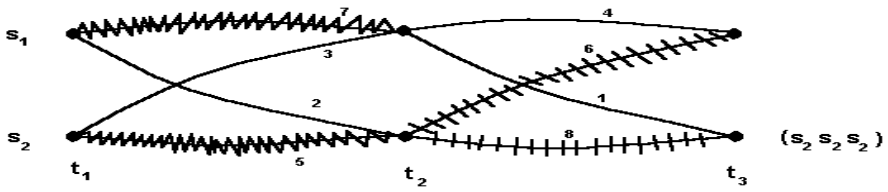
=> erhebliche Vereinfachung der Berechnung.

Ermittlung des besten verborgenen Pfades (1)

2. Frage : **Gegeben eine beobachtete Folge und ein Modell, welches ist der wahrscheinlichste verborgene Pfad ?**

Antwort : **Viterbi-Alignment**

Beispiel : **2 Zustände**



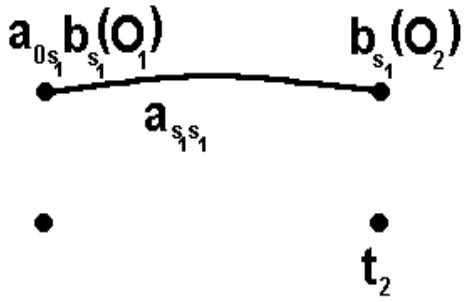
Ermittlung des besten verborgenen Pfades (2)

t_2 **bester Pfad muß durch s_1 oder s_2 gehen.**

- **wenn durch s_1, dann ist $s_1 s_1$ das optimale Anfangsstück,**
- **wenn durch s_2, dann ist $s_2 s_2$ das optimale Anfangsstück**

=> Für jeden Zeitpunkt t ist das Anfangsstück des insgesamt besten Pfades einer der bis t besten Teilpfade für jeden Zustand $s \in S$.

Ermittlung des besten verborgenen Pfades (3)



Wenn s_1 der beste Vorgänger von s_1 ist, dann gilt hier

$$\delta_2(s_1) = a_{0s_1} b_{s_1}(O_1)$$

Wahrscheinlichkeit des Pfades und der Ausgaben bis t=2.

Ermittlung des besten verborgenen Pfades (4)

Viterbi Alignment

<u>Idee</u>: Unter den zu einem Zeitpunkt für jeden Zustand besten Pfaden muß das Anfangsstück des insgesamt besten Pfades sein.

Die Viterbi-Aufreihung (1)

<u>Algorithmus</u>: (Notation $|S|=N$: Es gibt N Zustände)

<u>Schritt 1</u> : $\delta_1(s_i) = a_{0s_i} b_{s_i}(O_1)$

[ist dasselbe wie $\alpha_1(s_i)$ **]**

("Ertrag" für den Pfad)

$\Psi_1(s_i) = 0$

(bestmögliche Vorgänger speichern)

Die Viterbi-Aufreihung (2)

<u>Schritt 2</u> : Rekursion über Zeitpunkte und Zustände

$$\delta_t(s_j) \quad = \quad \max_{s_i \in S}\left(\delta_{t-1}(s_i) a_{s_i s_j}\right) \cdot b_{s_j}(O_t)$$

bestmöglicher Übergang von $t-1$ nach t zu s_j

(maximaler Ertrag für s_j + Übergang zu s_i mal Ausgabewahrscheinlichkeit O_t in s_j;

entspricht Variable V bei Huang et al., Kap. 8)

$$\Psi_t(s_j) = \arg\max_{s_i \in S}\left(\delta_{t-1}(s_i) a_{s_i s_j}\right)$$

"dasjenige s_i, für das (. . .) am größten wird"

Die Viterbi-Aufreihung (3)

Schritt 3 : Beendigung t=T :

$$P^* = \max_{s_i \in S} \delta_t(s_i)$$

$$I_T^* = \arg\max_{s_i \in S}(\delta_t(s_i))$$

I_T^* : **der Endpunkt(-zustand) des besten Pfades**

P^* : **die Pfadwahrscheinlichkeit**

Die Viterbi-Aufreihung (4)

Schritt 4 : Rückverfolgung des besten Pfades :

für t=T-1 bis 1 :

$$I_t^* = \Psi_{t+1}(I_{t+1}^*)$$

Zustand des besten Pfades in t

Schätzung der Modellparameter (1)

3. Frage : Gegeben beobachtete Folge, wie können die Parameter gesetzt werden, um diese Folge maximal wahrscheinlich zu machen ?

Problem: Wir beobachten Output-Sequenzen := unvollständige Daten (da wir den erzeugenden verborgenen Pfad nicht kennen). Um die Parameter der Übergangs- und Ausgabe-Wahrscheinlichkeiten zu schätzen, müssen wir auf die „plausibelsten" vollständigen Daten (Output-Sequenzen bezogen auf verborgene Pfade) zurück schließen.

Lösungsansatz: EM-Algorithmus (Dempster, A., Laird, N., Rubin, D. (1977): Maximum Likelihood from

Incomplete Data via the EM Algorithm. J. Roy. Stat. Soc., B, 39, pp. 1 - 38.):

Wir beobachten unvollständige Daten und schätzen die Wahrscheinlichkeiten vollständiger Daten (E-rwartungswertbildung vollständiger Daten).

Dann ändern wir die Parameter des HMM so ab, daß diese Erwartungswerte maximal wahrscheinlich werden (M-aximierung).

In der Praxis wird diese Maximierung bereits bei der maximum-likelihood Schätzung der vollständigen Daten erreicht. Daß dies so ist, weist Jelinek, Kapitel 9 nach.

Bei HMM mit kontinuierlichen Outputs (ASR: Konvexkombinationen von Normalverteilungen) werden entsprechend die Parameter der Normalverteilungen und die Mischungskoeffizienten geschätzt (s. Huang et al., Kap. 8.3).

Umsetzung EM für HMM setzt 3 Definitionen voraus .

1. "backward-Variable"(rückwärts die Kette verfolgend)

$\beta_t(s_i) = P_\lambda(O_{t+1}...O_T | s_t = s_i)$

$\beta_T(s) = 1$ für alle $s \in S$

(Folgerung : $P_\lambda(O, s_t = s) = \alpha_t(s)\beta_t(s)$

$= P_\lambda(O_1...O_t, s_t = s) P_\lambda(O_{t+1}...O_T | s_t = s)$ **)**

Schätzung der Modellparameter (2)

2. $\gamma_t(s) = P_\lambda(s_t = s | O) = P_\lambda(s_t = s, O) / P_\lambda(O)$

[Wahrscheinlichkeit, zum Zeitpunkt *t* in *s* zu sein, gegeben beobachtete Folge]

$$= \frac{\alpha_t(s)\beta_t(s)}{P_\lambda(O)}$$

Summe $\sum_{t=1}^{T} \gamma_t(s)$ **gibt an, mit welcher Wahrscheinlichkeit Zustand s insgesamt angenommen wird.**

Dies ist gerade die gesuchte Schätzung für die Wahrscheinlichkeiten der nicht beobachteten latenten Zustände des HMM.

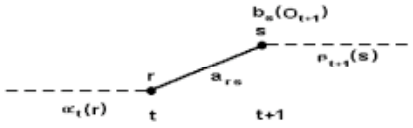
Schätzung der Modellparameter (3)

3. $\gamma_t(r,s) = P_\lambda(s_t = r, s_{t+1} = s | O)$

[Wahrscheinlichkeit, von t nach t+1 den Übergang vom Zustand r in den Zustand s zu machen (gegebene Beobachtungen)]

$$= P_\lambda(s_t = r, s_{t+1} = s, O) / P_\lambda(O)$$

$$= \alpha_t(r) a_{rs} b_s(O_{t+1}) \beta_{t+1}(s) / P_\lambda(O)$$



Dies ist tger ade die gesuchte Schätzung für die Wahrscheinlichkeiten der nicht beobachteten Übergänge zwischen latenten Zuständen des HMM.

Schätzung der Modellparameter (4)

Ausführliche Herleitung:

$$P_\lambda(s_t = r, s_{t+1} = s, O) / P_\lambda(O)$$
$$= P_\lambda(O_1 ... O_t, s = r) P_\lambda(s_{t+1} = s | s_t = r)$$
$$P_\lambda(O_{t+1} ... O_T | s_{t+1} = s) / P_\lambda(O)$$

Nun ist

$$P_\lambda(O_{t+1} ... O_T | s_{t+1} = s) = P_\lambda(O_{t+1} | s_{t+1} = s) P_\lambda(O_{t+2} ... O_T | s_{t+1} = s)$$

Also ist

$$P_\lambda(s_t = r, s_{t+1} = s, O) / P_\lambda(O)$$
$$= \alpha_t(r) a_{rs} b_s(O_{t+1}) \beta_{t+1}(s) / P_\lambda(O)$$

Schätzung der Modellparameter (5)

<u>Überlegung</u> : Wenn Parameter so abgeändert werden, daß die angenommenen Wahrscheinlichkeiten für Zustände bzw. Übergänge den beobachteten Häufigkeiten entsprechen, dann muß sich die Wahrscheinlichkeit $P_\lambda(O)$ vergrößern, für neues

$$\lambda' : P_{\lambda'}(O) \geq P_\lambda(O).$$

<u>Umsetzung</u> : ("Baum / Welch reestimates")

- $a'_{0 s_i} = \gamma_1(s_i)$ **(neue Anfangswahrscheinlichkeit für s_i := Wahrscheinlichkeit unter O mit s_i einen Pfad zu beginnen).**

Schätzung der Modellparameter (6)

- $$a'_{rs} = \sum_{t=1}^{T-1} \gamma_t(r,s) / \sum_{t=1}^{T-1} \sum_{s' \in S} \gamma_t(r,s')$$

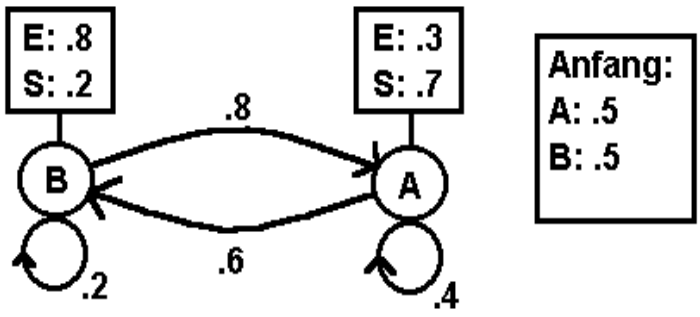
  (neue Übergangswerte für r → s ist relative geschätzte Häufigkeit des Übergangs r → s bezogen auf alle anderen s'.)

- $$b'_s(o) = \sum_{t:O_t = o} \gamma_t(s) / \sum_{t=1}^{T} \gamma_t(s)$$

  (neue Emissionswahrscheinlichkeiten: relative geschätzte Häufigkeit der Ausgabe o über alle Pfade zu allen Zeitpunkten.)

Ein Beispiel (1)

Beispiel :



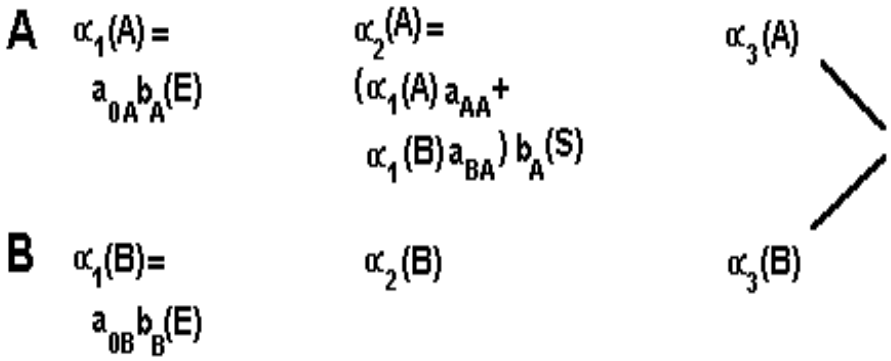
Ein Beispiel (2)

Beobachtung:

E S E.

1. Wie wahrscheinlich ist " E S E " ?

t_1	t_2	t_3
E	**S**	**E**

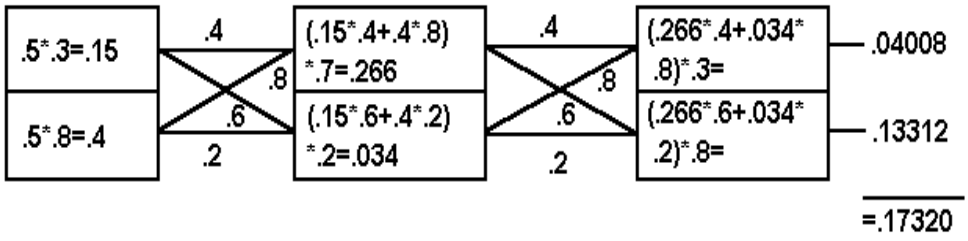


$$+ = P_\lambda(O)$$

Ein Beispiel (3)

t_1 t_2 t_3

E **S** **E**



A $\delta_1(A) = .15$ $\delta_2(A) = .32 * .7 = .224$ $.0896 * .3 = .02688$

$\Psi_2(A) = B$ $\Psi_3(A) = A$

B	$\delta_1(B) = .4$	$\delta_2(B) = .09 * .2 = .018$	$0.1344 * .8 = .10752$
		$\Psi_2(B) = A$	$\Psi_3(B) = A$

Ein Beispiel (4)

Wahrscheinlichstes Pfadende : B mit $P_T^* = .10752$

(das sind mehr als 60% der Gesamtwahrscheinlichkeit dieser Beobachtungsfolge)

Backtracking:

$$I_3^* = B, I_2^* = \Psi_3(B) = A, I_1^* = \Psi_2(A) = B$$

bester Pfad : B A B.

Ausblick: Von latenten Pfaden zu latenten Strukturen

Stochastische kontextfreie Grammatiken :