Hidden Markov Models (HMM)

Kursfolien

Karin Haenelt

Themen

- Definitionen
  - Stochastischer Prozess
  - Markow Kette
  - (Visible) Markov Model
  - Hidden Markov Model
- Aufgaben, die mit HMMs bearbeitet werden
- Algorithmen
  - Viterbi-Algorithmus
- Formen vom Hidden Markov Models
  - state emission model / arc emission model
  - ergodic model

Hidden Markov Model

- Hidden Markov Models (HMM)
  sind stochastische Modelle, die auf Markow-Ketten beruhen

Stochastischer Prozess

Definition 1

Ein stochstischer Prozess oder Zufallsprozess ist eine Folge von elementaren Zufallsereignissen

$$X_1, X_2, \ldots, X_i \in \Omega, i = 1,2,\ldots$$

Definition 2

Die möglichen Zufallswerte in einem stochastischen Prozess heißen Zustände des Prozesses.

Man sagt, dass sich der Prozess zum Zeitpunkt t in Zustand X_t befindet.

Brants, 1999: 30

Stochastischer Prozess

Beispiel

- Ein Textgenerator hat ein Lexikon mit drei Wörtern
- von denen an jeder Position jedes auftreten kann
  ($\Omega = \{geschickt, werden, wir\}$)
- wir beobachten an jeder Position, welches Wort generiert wurde
- Sei
  - X_1 das Wort zum ersten Beobachtungszeitpunkt
  - X_2 das Wort zum zweiten Beobachtungszeitpunkt, usw.
- Dann ist die Folge der Wörter ein stochastischer Prozess mit diskreter Zufallsvariable und diskretem Zeitparameter

Stochastischer Prozess

- Für die vollständige Beschreibung eines Zufallsprozesses mit diskretem Zeitparameter benötigt man
  1. die Anfangswahrscheinlichkeit:
     die für jeden Zustand angibt, mit welcher Wahrscheinlichkeit er als Zustand X_1 beobachtet werden kann (d.h. den Startzustand bildet)
     $$\pi_i = P(X_1 = s_i)$$
  2. die Übergangswahrscheinlichkeit:
     die für jeden Zustand angibt, mit welcher Wahrscheinlichkeit er in einer Zustandsfolge auftritt:
     $$P(X_{t+1} = x_{t+1} \mid X_1 = x_1, X_2 = x_2, \ldots X_t = x_t)$$

Brants, 1999: 30

Markow-Kette

Definition 3

Eine Markow-Kette ist ein spezieller stochastischer Prozess,
bei dem zu jedem Zeitpunkt
die Wahrscheinlichkeiten aller zukünftigen Zustände
nur vom momentanen Zustand abhängen
(= Markow-Eigenschaft)

d.h. es gilt:

$$P(X_{t+1} = x_{t+1} \mid X_1 = x_1, X_2 = x_2, \ldots X_t = x_t) =$$
$$P(X_{t+1} = x_{t+1} \mid X_t = x_t)$$

Brants, 1999: 30

endliche Markow-Kette

Definition 4

Für eine endliche Markow-Kette gibt es endlich viele Zustände, und die Kette muss sich zu jedem Zeitpunkt in einem dieser endlich vielen Zustände befinden

Brants, 1999: 31

Prozess „ohne Gedächtnis“ (Def 3)
mit endlich vielen Zuständen (Def 4),
entspricht den Eigenschaften eines endlichen Automaten

Markow-Kette: Matrix-Darstellung

kann beschrieben werden durch die Angaben

- Stochastische Übergangsmatrix A

$$a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$$

$$\forall i, j \qquad a_{ij} \geq 0$$

$$\forall i \qquad \sum_{j=1}^{N} a_{i,j} = 1$$

$X_t = s_i$	$X_{t+1} = s_j$		
	geschickt	werden	wir
geschickt	.3	.4	.3
werden	.4	.2	.4
wir	.3	.4	.3

- Anfangswahrscheinlichkeiten Π

$$\pi_i = P(X_1 = s_i)$$

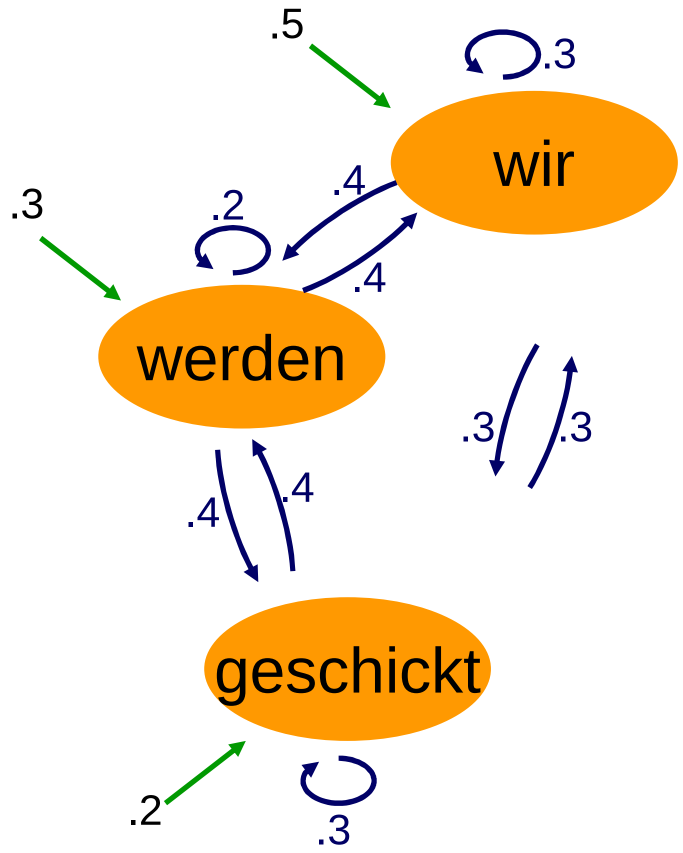
$$\sum_{i=1}^{N} \pi_i = 1$$

X t	π
geschickt	.2
werden	.3
wir	.5

Manning/Schütze, 2000: 318

Markow-Kette: Graph-Darstellung

kann beschrieben werden durch einen Zustandsübergangsgraphen



Markow-Kette: Berechnung einer Sequenz-Wahrscheinlichkeit

Wahrscheinlichkeit der Sequenz der Zustände X_1 … X_T

$$P(X_1,\ldots,X_T)$$

$$= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2, X_1)\ldots P(X_T \mid X_1,\ldots,X_{T-1})$$

für eine Markow-Kette gilt:

$$= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)\ldots P(X_T \mid X_{T-1})$$

$$= \pi_{X_1} \prod_{t=1}^{T-1} a_{X_t X_{t+1}}$$

Manning/Schütze, 2000: 320

Markow-Kette: Berechnungsbeispiel

Wahrscheinlichkeit der Sequenz der Zustände $X_1 \dots X_T$

$$P(X_1 = wir, X_2 = werden, X_3 = geschickt)$$

$$= P(X_1 = wir) \cdot$$
$$P(X_2 = werden \mid X_1 = wir) \cdot$$
$$P(X_3 = geschickt \mid X_2 = werden)$$
$$= (.5 \times .4 \times .4) = 0.08$$

X t	π
geschickt	.2
werden	.3
wir	.5

$X_t = s_i$	$X_{t+1} = s_j$		
	geschickt	werden	wir
geschickt	.3	.4	.3
werden	.4	.2	.4
wir	.3	.4	.3

Markow-Modell (MM)

- Ein Markow-Modell ordnet jedem Zustand (andere Variante: jedem Zustandsübergang) eine Ausgabe zu, die ausschließlich vom aktuellen Zustand (bzw. Zustandsübergang) abhängig ist
- Ausgabe: Sequenz von Ereignissen, die die Beobachtungen in der Beobachtungssequenz repräsentieren

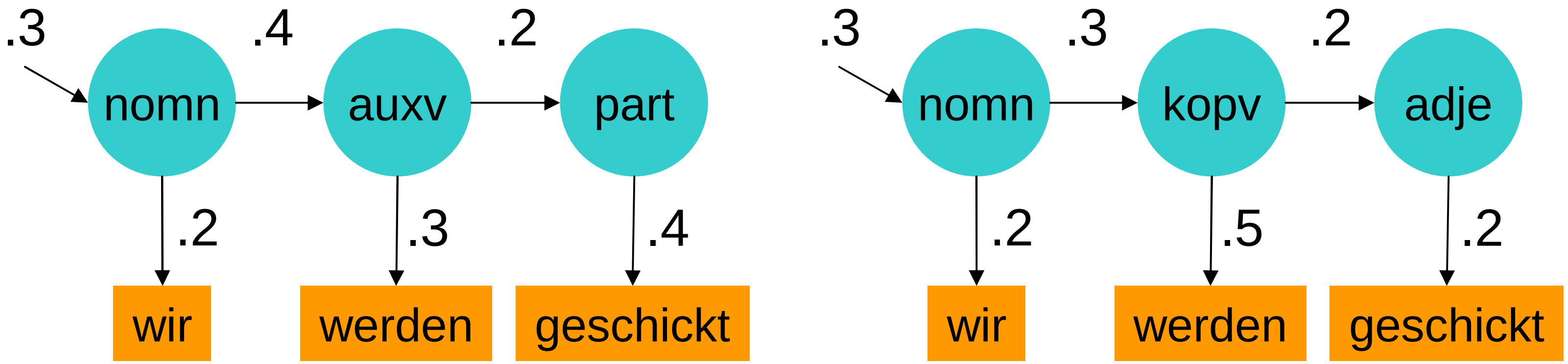
Zur Unterscheidung auch Visible Markov Model (VMM) genannt

Hidden Markov Modell (HMM): Beschreibung

- Ein Hidden Markov Model ist ein Markow-Modell
  - bei dem nur die Sequenz der Ausgaben beobachtbar ist,
  - die Sequenz der Zustände verborgen bleibt
- Es kann mehrere Zustandssequenzen geben, die dieselbe Ausgabe erzeugen

Hidden Markov Modell: Beispiel

- in einem Text lassen sich nur die Ausgaben (= produzierte Wörter) beobachten (visible)
- die Sequenz von Zuständen (= Wortarten), die die Wörter ausgeben, (Satzmuster) lässt sich nicht beobachten (hidden)
- mehrere Sequenzen können dieselbe Ausgabe erzeugen:



.3 x .2 x .4 x .3 x .2 x .4 =0.000576

.3 x .2 x .3 x .5 x .2 x .2 =0.000360

Hidden Markov Model: Definition

Ein HMM wird spezifiziert durch ein Fünf-Tupel (S,K, Π, A, B)			
$S = \{S_1, ..., S_N\}$	Menge der Zustände		
$K = \{k_1, ..., k_M\}$	Menge der Ausgabesymbole		
$\Pi = \{\pi_i\}$	Wahrscheinlichkeiten der Startzustände		
	$\pi_i = P(X_1 = S_i)$		$\sum_{i=1}^{N} \pi_i = 1$
$A = \{a_{ij}\}$	Wahrscheinlichkeiten der Zustandsübergänge		
	$a_{ij} = P(X_{t+1} = S_j \mid X_t = S_i)$	$1 \leq i$, $j \leq N$	$\sum_{j=1}^{N} a_{ij} = 1$
$B = \{b_j(k)\}$	Wahrscheinlichkeiten der Symbolemissionen in Zustand j		
	$b_j(k) = P(K_k \text{ in } t \mid X_t = S_j)$	$1 \leq j \leq N$ $1 \leq k \leq M$	$\sum_{k=1}^{M} b_j(k) = 1$

Rabiner, 1989, S. 260/261

Manning/Schütze, 2000: 318-324

Ein Hidden Markov Model

	Übergangsmatrix					Emissionsmatrix				Startwahrscheinlichkeit
X_t	X_{t+1}					o_t				π
	Adje	AuxV	KopV	Nomn	Part	geschickt	werden	wir	...	
Adje	.2	.1	.1	.4	.2	.2	0	0	.8	.3
AuxV	.2	.3	.1	.2	.2	0	.3	0	.7	.2
KopV	.2	.2	.1	.4	.1	0	.5	0	.5	.1
Nomn	.1	.4	.3	.1	.1	0	0	.2	.8	.3
Part	.3	.1	.2	.1	.3	.4	0	0	.6	.1

Hidden Markov Model: Gewinnung der Daten – Übersicht

- Annotation eines Corpus
- Auszählung der Sequenzen
- Umrechnung der Häufigkeiten in prozentuale Anteile

Hidden Markov Model: Gewinnung der Daten (1)

- Annotation eines Corpus
- Auszählung der Sequenzen
- Umrechnung der Häufigkeiten in prozentuale Anteile

wir	werden	geschickt	vom	König	.
nomn	auxv	part	..	..	Ω

Wir	werden	geschickt	durch	Übung	.
nomn	kopv	adje	..	…	Ω

Hidden Markov Model: Gewinnung der Daten (2)

- Annotation eines Corpus
- Auszählung der Sequenzen
- Umrechnung der Häufigkeiten in prozentuale Anteile

	Adje	AuxV	KopV	Nomn	Part	Ω	geschickt	werden	wir	.
Adje	-	-	-	-	-	1	1	-	-	-
AuxV	-	-	-	-	1	-	-	1	-	-
KopV	1	-	-	-	-	-	1	-	-	-
Nomn	-	1	1	-	-	-	-	-	2	-
Part	-	-	-	-	-	1	-	-	-	-
Ω	-	-		1	-	-	-	-	-	2

Hidden Markov Model: Gewinnung der Daten (3)

- Annotation eines Corpus
- Auszählung der Sequenzen
- Umrechnung der Häufigkeiten in prozentuale Anteile

	Adje	AuxV	KopV	Nomn	Part	Ω	geschickt	werden	wir	.
Adje	-	-	-	-	-	1.0	1.0	-	-	-
AuxV	-	-	-	-	1.0	-	-	1.0	-	-
KopV	1.0	-	-	-	-	-	1.0	-	-	-
Nomn	-	0.5	0.5	-	-	-	-	-	1.0	-
Part	-	-	-	-	-	1.0	-	-	-	-
Ω	-	-		1.0	-	-	-	-	-	1.0

Drei grundlegende Aufgaben, die mit HMMs bearbeitet werden

1. Dekodierung: Wahrscheinlichkeit einer Beobachtung finden
   - brute force
   - Forward-Algorithmus / Backward-Algorithmus
2. Beste Pfad-Sequenz finden
   - brute force
   - Viterbi-Algorithmus
3. Training: Aufbau des besten Modells aus Trainingsdaten

Manning/Schütze, 2000: 325

A1: Wahrscheinlichkeit einer Beobachtung finden

gegeben eine Sequenz von Beobachtungen $O = (o_1, \ldots, o_T)$

O=(*wir*,*werden*,*geschickt*)

ein Modell $\mu = (A, B, \Pi)$

	Adje	AuxV	KopV	Nomn	Part	g'schickt	werden	wir	..	π
Adje	.2	.1	.1	.4	.2	.2	0	0	.8	.3
AuxV	.2	.3	.1	.2	.2	0	.3	0	.7	.2
KopV	.2	.2	.1	.4	.1	0	.5	0	.5	.1
Nomn	.1	.4	.3	.1	.1	0	0	.2	.8	.3
Part	.3	.1	.2	.1	.3	.4	0	0	.6	.1

gesucht die Wahrscheinlichkeit $P(O \mid \mu)$

$$P(wir, werden, geschickt \mid \mu)$$

A1: Wahrscheinlichkeit einer Beobachtung finden

<u>Lösungsweg 1: brute force</u>

Für alle möglichen Zustandsfolgen

- Berechnung der Wahrscheinlichkeit der Beobachtungen
- Summierung der Wahrscheinlichkeiten

$$P(O \mid \mu) = \sum_{X} P(O \mid X, \mu) P(X \mid \mu)$$

$$= \sum_{X_1 \ldots X_T} \pi_{X_1} b_{X_1 O_1} \prod_{t=1}^{T-1} \underbrace{a_{X_t X_{t+1}}}_{\text{state transition}} \underbrace{b_{X_{t+1} O_{t+1}}}_{\text{symbol emission}}$$

vgl. Rabiner, 1989, S. 260/261
vgl. Manning/Schütze, 2000: 326

A1: Wahrscheinlichkeit einer Beobachtung finden

<u>Lösungsweg 1: brute force: Beispiel</u>

$$P(O|\mu) \quad = \sum_{X_1 \ldots X_T} \pi_{X_1} b_{X_1 O_1} \prod_{t=1}^{T-1} a_{X_t X_{t+1}} b_{X_{t+1} O_{t+1}}$$

P(*wir,werden,geschickt* | Adje Adje Adje, μ)		=0.0
+ P(*wir,werden,geschickt* | Adje Adje AuxV, μ)		
+ …		
+ P(*wir,werden,geschickt* | Nomn AuxV Part, μ)	.3 x .2 x .4 x .3 x .2 x .4	=0.000576
+ …		
+ P(*wir,werden,geschickt* | Nomn KopV Adje, μ)	.3 x .2 x .3 x .5 x .2 x .2	=0.000360
+ …		
+ P(*wir,werden,geschickt* | Part Part Part, μ)		=0.0
= …		=0.000936

A1: Wahrscheinlichkeit einer Beobachtung finden

<u>Lösungsweg 1: brute force: Effizienz</u>

$$P(O|\mu) = \sum_{X_1 \ldots X_T} \pi_{X_1} b_{X_1 O_1} \prod_{t=1}^{T-1} a_{X_t X_{t+1}} b_{X_{t+1} O_{t+1}}$$

Lösungsweg ist <u>hoffnungslos ineffizient</u>

Benötigt im allgemeinen Fall, d.h.

- Start in jedem Zustand möglich,
- Jeder Zustand kann auf jeden folgen

$(2T - 1) \times N^T$ Multiplikationen

T Anzahl der Beobachtungen

N Anzahl der Zustände

vgl. Rabiner, 1989, S. 260/261

vgl. Manning/Schütze, 2000: 326

A1: Wahrscheinlichkeit einer Beobachtung finden

Lösungsweg 2: Vorwärts- und Rückwärts-Verfahren

Forward procedure
Backward procedure

Merken partieller Ergebnisse statt
Wiederholter Berechnung

Manning/Schütze, 2000: 326ff

A2: Beste Pfadsequenz finden

gegeben eine Sequenz von Beobachtungen $O = (o_1, \ldots, o_T)$

O=(*wir*,*werden*,*geschickt*)

ein Modell $\mu = (A, B, \Pi)$

	Adje	AuxV	KopV	Nomn	Part	g'schickt	werden	wir	π
Adje	.2	.1	.1	.4	.2	.2	0	0	.3
AuxV	.2	.2	.2	.2	.2	0	.3	0	.2
KopV	.2	.2	.2	.3	.1	0	.5	0	.1
Nomn	.1	.4	.3	.1	.1	0	0	.2	.3
Part	.3	.1	.2	.1	.3	.4	0	0	.1

gesucht die wahrscheinlichste Pfadsequenz

$P(wir, werden, geschickt \mid \mu)$ $\arg_X \max P(X \mid O, \mu)$

A2: Beste Pfadsequenz finden

<u>Lösungsweg 1: brute force:</u>

Wie in [A1]: alle Varianten berechnen
die wahrscheinlichste auswählen
hoffnungslos ineffizient

<u>Lösungsweg 2: beste Einzelzustände</u>

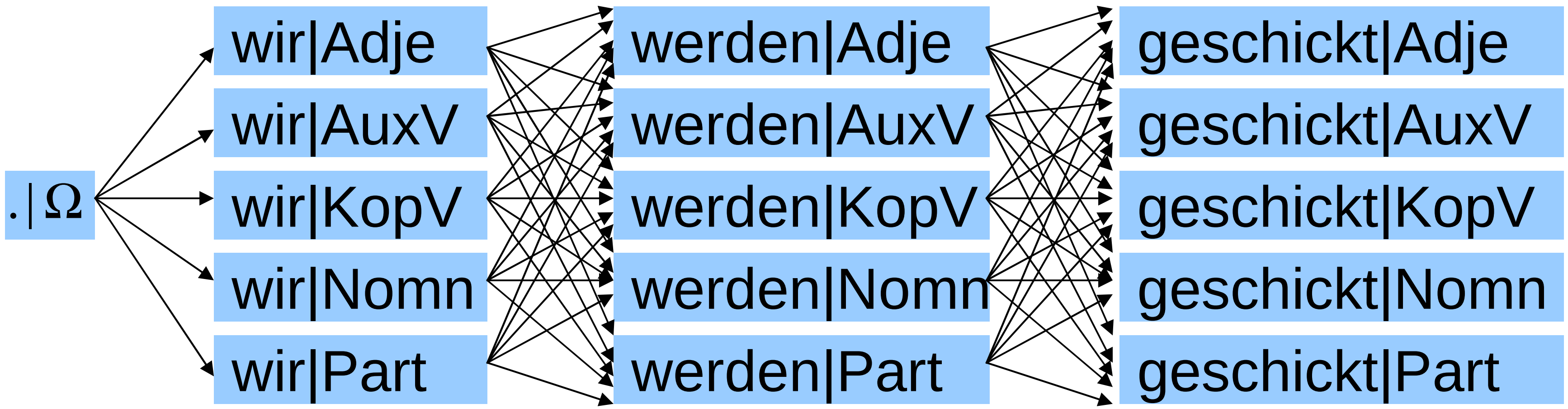
Für jeden Zeitpunkt t
Zustand mit höchster Ausgabewahrscheinlichkeit auswählen
Zusammensetzung kann unwahrscheinliche Sequenzen ergeben

A2: Beste Pfadsequenz finden

<u>Lösungsweg 3: Viterbi-Algorithmus</u>

Speichert für jeden Zeitpunkt t

die Wahrscheinlichkeit des wahrscheinlichsten Pfades, der zu einem Knoten führt



A3: Training der Modellparameter

gegeben <u>eine Sequenz von Beobachtungen</u> $O = (o_1, \ldots, o_T)$
<u>In einem Trainingscorpus</u>

gesucht <u>ein Modell</u> $\mu = (A, B, \Pi)$

das für die beobachteten Sequenzen im Trainingscorpus
die maximalen Wahrscheinlichkeiten erzeugt

$$\arg_{\mu} \max P(O_{Training} \mid \mu)$$

Manning/Schütze, 2000: 333ff

A3: Training der Modellparameter

Lösung Baum-Welch oder Forward-backward-Algorithmus

Manning/Schütze, 2000: 333ff

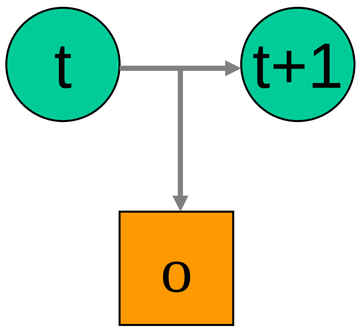
Formen von Hidden Markov Models: Emissionen

- auf den vorangehenden Folien wurde ein State Emission Model verwendet
- den allgemeinen Fall stellt ein Arc Emission Model dar
- ein State Emission Model kann in ein Arc Emission Model überführt werden, umgekehrt ist dies nicht immer möglich

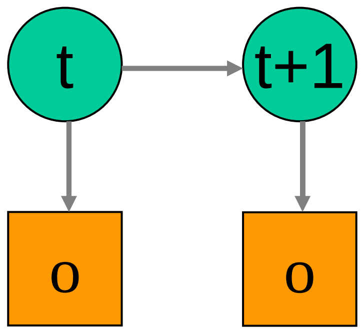
- auf den folgenden Folien wird ein Arc Emission Model beschrieben

Formen von Hidden Markov Models: Emissionen

- <u>Allgemeine Form:</u>
  Arc Emission Model
  - Zur Zeit t emittiertes Symbol
    hängt ab von
    - Zustand zur Zeit t und
    - Zustand zur Zeit t+1

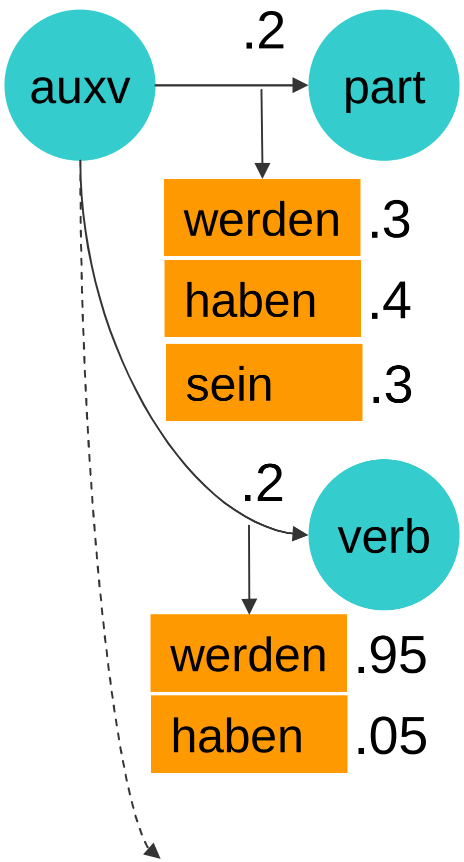


- <u>Spezielle Form:</u>
  State Emission Model
  - Zur Zeit t emittiertes Symbol
    hängt ab von
    - Zustand zur Zeit t

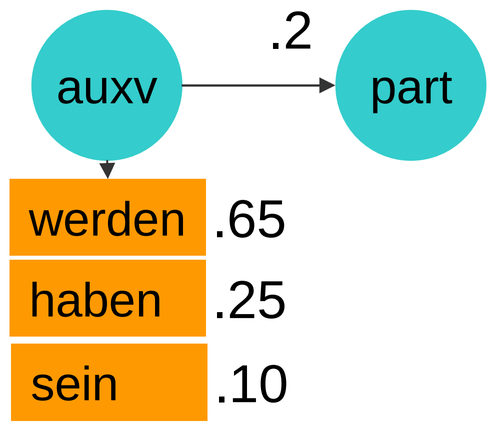


Formen von HMM: Emissionen: Beispiel

- Arc Emission Model

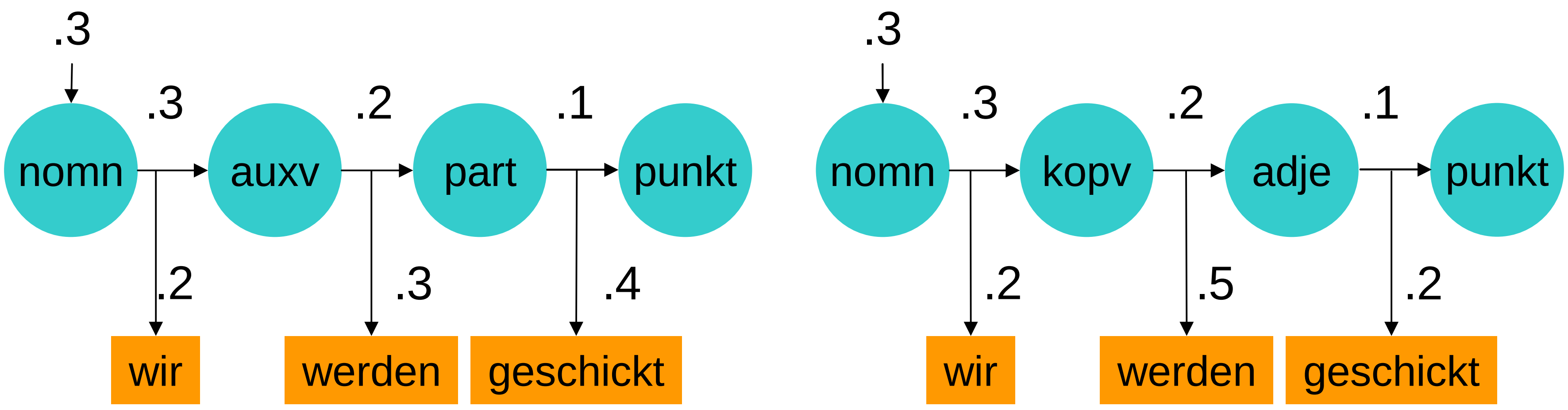


- State Emission Model



Arc Emission Model: Beispiel

- in einem Text lassen sich nur die Ausgaben (= produzierte Wörter) beobachten (visible)
- die Sequenz von Zuständen (= Wortarten), die die Wörter ausgeben, (Satzmuster) lässt sich nicht beobachten (hidden)
- mehrere Sequenzen können dieselbe Ausgabe erzeugen:



.3 x .3 x .2 x .2 x .3 x .1 x .4 = 0.0000432

.3 x .3 x .2 x .2 x .5 x .1 x .2 = 0.000036

Arc Emission Model:

Darstellung als Wahrscheinlichkeitsmatrix

	Übergangsmatrix						Start
X_t	X_{t+1}						
	Adje	AuxV	KopV	Nomn	Part	Punkt	π
Adje	.2	.1	.1	.4	.1	.1	.3
AuxV	.2	.3	.1	.1	.2	.1	.2
KopV	.2	.1	.1	.4	.1	.1	.1
Nomn	.05	.4	.3	.05	.1	.1	.3
Part	.3	.1	.1	.1	.3	.1	.1
Punkt	.2	.2	.1	.3	.1	.1	.1

Adje → Adje:

Emissionsmatrix			
o_t			
geschickt	werden	wir	...
.2	0	0	.8

KopV → Adje:

Emissionsmatrix			
o_t			
geschickt	werden	wir	...
0.05	.5	.05	.4

Arc Emission Model:

Spezialfall: State Emission Model

<table>
<tr><td></td><td colspan="2">Übergangsmatrix</td></tr>
<tr><td>X_t</td><td colspan="2">X_{t+1}</td></tr>
<tr><td></td><td>Adje</td><td>AuxV</td></tr>
<tr><td>Adje</td><td>.2
<table>
<tr><td colspan="4">Emissionsmatrix</td></tr>
<tr><td colspan="4">o_t</td></tr>
<tr><td>geschickt</td><td>werden</td><td>wir</td><td>...</td></tr>
<tr><td>.2</td><td>0</td><td>0</td><td>.8</td></tr>
</table></td><td>.2
<table>
<tr><td colspan="4">Emissionsmatrix</td></tr>
<tr><td colspan="4">o_t</td></tr>
<tr><td>geschickt</td><td>werden</td><td>wir</td><td>...</td></tr>
<tr><td>.2</td><td>0</td><td>0</td><td>.8</td></tr>
</table></td></tr>
<tr><td>AuxV</td><td></td><td></td></tr>
<tr><td>...</td><td></td><td></td></tr>
</table>

Wenn die Emissionsverteilungen
für alle Übergänge aus einem Zustand
identisch sind,
entspricht dies einem State Emission Modell

Arc Emission Model: Definition

Ein HMM wird spezifiziert durch ein Fünf-Tupel (S,K, Π, A, B)			
$S = \{S_1, ..., S_N\}$	Menge der Zustände		
$K = \{k_1, ..., k_M\}$	Menge der Ausgabesymbole		
$\Pi = \{\pi_i\}$	Wahrscheinlichkeiten der Startzustände		
	$\pi_i = P(X_1 = S_i)$		$\sum_{i=1}^{N} \pi_i = 1$
$A = \{a_{ij}\}$	Wahrscheinlichkeiten der Zustandsübergänge		
	$a_{ij} = P(X_{t+1} = S_j \mid X_t = S_i)$	$1 \leq i$, $j \leq N$	$\sum_{j=1}^{N} a_{ij} = 1$
$B = \{b_{ijk}\}$	Wahrscheinlichkeiten der Symbolemissionen		
	$b_{ijk} = P(K_k$ bei Übergang von X_t zu $X_{t+1} \mid X_t = S_j, X_{t+1} = S_j)$	$1 \leq j \leq N$ $1 \leq k \leq M$	$\sum_{k=1}^{M} bijk = 1$

Manning/Schütze, 2000: 318-324

A1: Wahrscheinlichkeit einer Beobachtung finden

<u>Lösungsweg 1: brute force</u>

Für alle möglichen Zustandsfolgen

- Berechnung der Wahrscheinlichkeit der Beobachtungen
- Summierung der Wahrscheinlichkeiten

$$P(O \mid \mu) = \sum_X P(O \mid X, \mu) P(X \mid \mu)$$

$$= \sum_{X_1 \ldots X_{t+1}} \pi_{X_1} \prod_{t=1}^{T} a_{X_t X_{t+1}} b_{X_t X_{t+1} o_t}$$

state transition

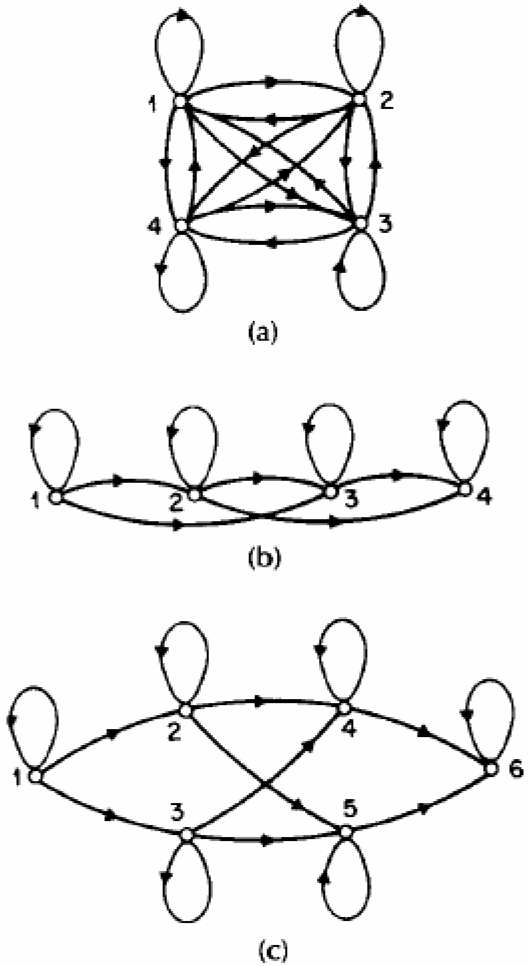
symbol emission

Manning/Schütze, 2000: 326

Formen von Hidden Markov Models: Verbindungen zwischen Zuständen

- ergodic model: jeder Zustand kann von jedem in einer endlichen Anzahl von Schritten erreicht werden:

- andere Arten z.B. in der Verarbeitung gesprochener Sprache verwendet



Rabiner, 1989, S. 266

Vielen Dank

Für das Aufspüren von Fehlern in früheren Versionen und Hinweise zur Verbesserung danke ich

Wiebke Petersen

versions: HMM 01.02.2005, 07.11.2004, 12.06.2004, 09.05.2002

Literatur

- Allen, James (1995): *Natural Language Understanding. 2nd edition. Addison-Wesley Publishing Co.*
- Brants, Thorsten (1999). *Statistische Methoden in der Sprachverarbeitung*. Seminarskript 15. Juni 1999
- Haenelt, Karin: *Der Viterbi-Algorithmus. Eine Erläuterung der formalen Spezifikation am Beispiel des Part-of-Speech Tagging.* Kursskript. 11.05.2002 http://kontext.fraunhofer.de/haenelt/kurs/folien/Viterbi-Tutor.doc http://kontext.fraunhofer.de/haenelt/kurs/folien/Viterbi-Tutor.htm
- Manning, Christopher D.; Schütze, Hinrich (1999): *Foundations of Statistical Natural Language Processing.* Cambridge, Mass., London: The MIT Press. (vgl.: http://www.sultry.arts.usyd.edu.au/fsnlp)
- Rabiner, Lawrence R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: *Proceedings of the IEEE*, Vol. 77, No. 2, February. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf