Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung II / 11 FS 2017

Statistische Sprachmodellierung

Buch: Kapitel 14.1 und 14.2

Beat Pfister

Programm heute

Vorlesung: Sprachmodellierung (Teil 1)

- Was ist Sprachmodellierung?
- Statistischer Ansatz der Sprachmodellierung

Übung: statistische Sprachmodellierung

Sprachmodellierung in der Spracherkennung

- Sprachmodell: Sammlung von A-priori-Kenntnissen über die Sprache
  Vorwissen, Erwartung
  (*speech modelling* vs. *language modelling*)
- Information: Sprache ist ein Informationsträger
  (etwas Neues, Unvorhergesehenes)

$\longrightarrow$ Sprachmodell und Information sind komplementär!

Ziel der Sprachmodellierung:

Dem Spracherkenner möglichst viel Vorwissen zu geben
um die Fehlerrate zu reduzieren.

Wissensbasierte vs. statistische Sprachmodelle

- Sprachmodell: besteht aus verschiedenen Teilmodellen wie Vokabular, Aussprachelexikon, Satzgrammatik ...
- Inhalt: linguistisches Wissen und Erfahrungswerte
- Formalismus: Darstellung richtet sich nach der Art des Wissens
  - wissensbasiert: Lexika, Regeln
  - statistisch: Wahrscheinlichkeit von Wörtern, Ausdrücken, Sätzen ...

Statistische Spracherkennung

gegeben: Vokabular V und Merkmalssequenz $\mathbf{X}$

gesucht: Wortfolge $\hat{W}$ aus V^* mit: $\hat{W} = \underset{W \in V^*}{\operatorname{argmax}} P(W|\mathbf{X})$ MAP-Regel

wobei: $P(W|\mathbf{X}) = \dfrac{P(\mathbf{X}|W) \cdot P(W)}{P(\mathbf{X})}$

$P(\mathbf{X})$	Wahrscheinlichkeit der Merkmalssequenz $\mathbf{X}$
$P(W)$	A-priori-Wahrscheinlichkeit der Wortfolge W $\longrightarrow$ statistisches Sprachmodell
$P(\mathbf{X}\|W)$	akustisches Modell (z.B. ein HMM), das beschreibt, mit welcher Wahrscheinlichkeit sich die Wortfolge W in der Merkmalssequenz $\mathbf{X}$ manifestiert

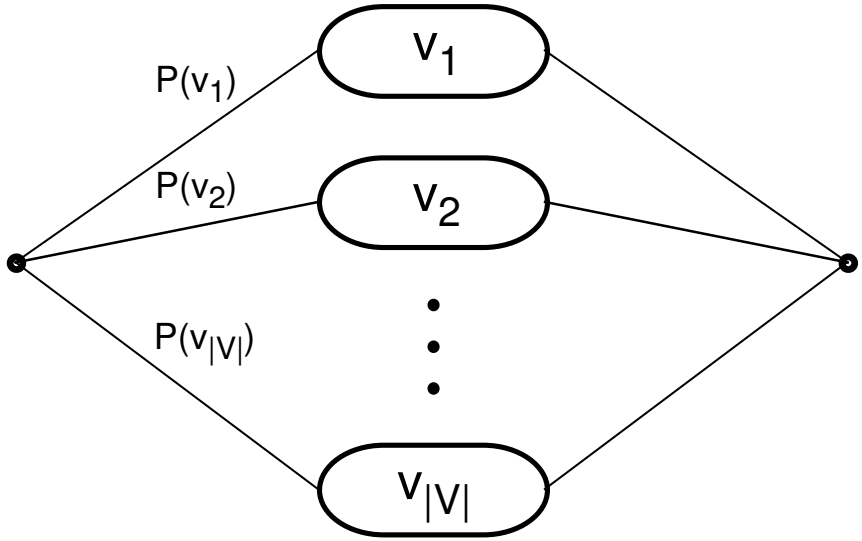
Sprachmodellierung bei Einzelworterkennung

Aufgabe: Erkennen eines Wortes w aus dem Vokabular $V = \{v_1, v_2, \ldots, v_{|V|}\}$

Sprachmodell: $P(v_i), \quad v_i \in V$ A-priori-Wahrscheinlichkeit der Wörter

Training: $$P(v_i) \approx \text{freq}(v_i) = \frac{\text{cnt}(v_i)}{|\text{Stichprobe}|}$$

Anwendung: Erkennungsnetzwerk mit Wortwahrscheinlichkeiten



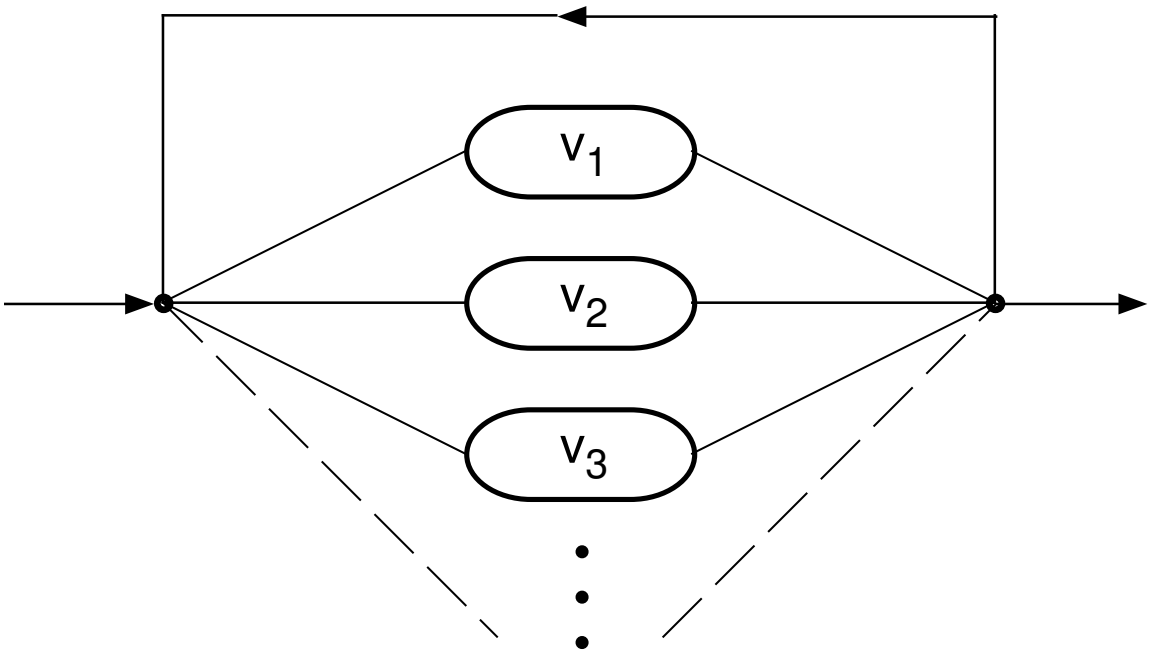
Bemerkung: $P(v_i)$ sind von der Anwendung abhängig!

Erkennen von Wortfolgen

Aufgabe: Erkennen einer zusammenhängenden Wortfolge $W_1^K = w_1 w_2 \ldots w_K$, wobei die Anzahl der Wörter unbekannt ist:

Ansatz: Erkennungsnetzwerk

"Word Loop"



>>>

Sprachmodellierung für Wortfolgen

Aufgabe: Erkennen einer logisch zusammenhängenden Wortfolge:
$W_1^K = w_1 w_2 \dots w_K$

Sprachmodell: muss Zusammenhang zwischen den Wörtern berücksichtigen:
$P(W_1^K) = P(w_1 \dots w_K) \quad \neq P(w_1) \cdot P(w_2) \cdots P(w_K)$

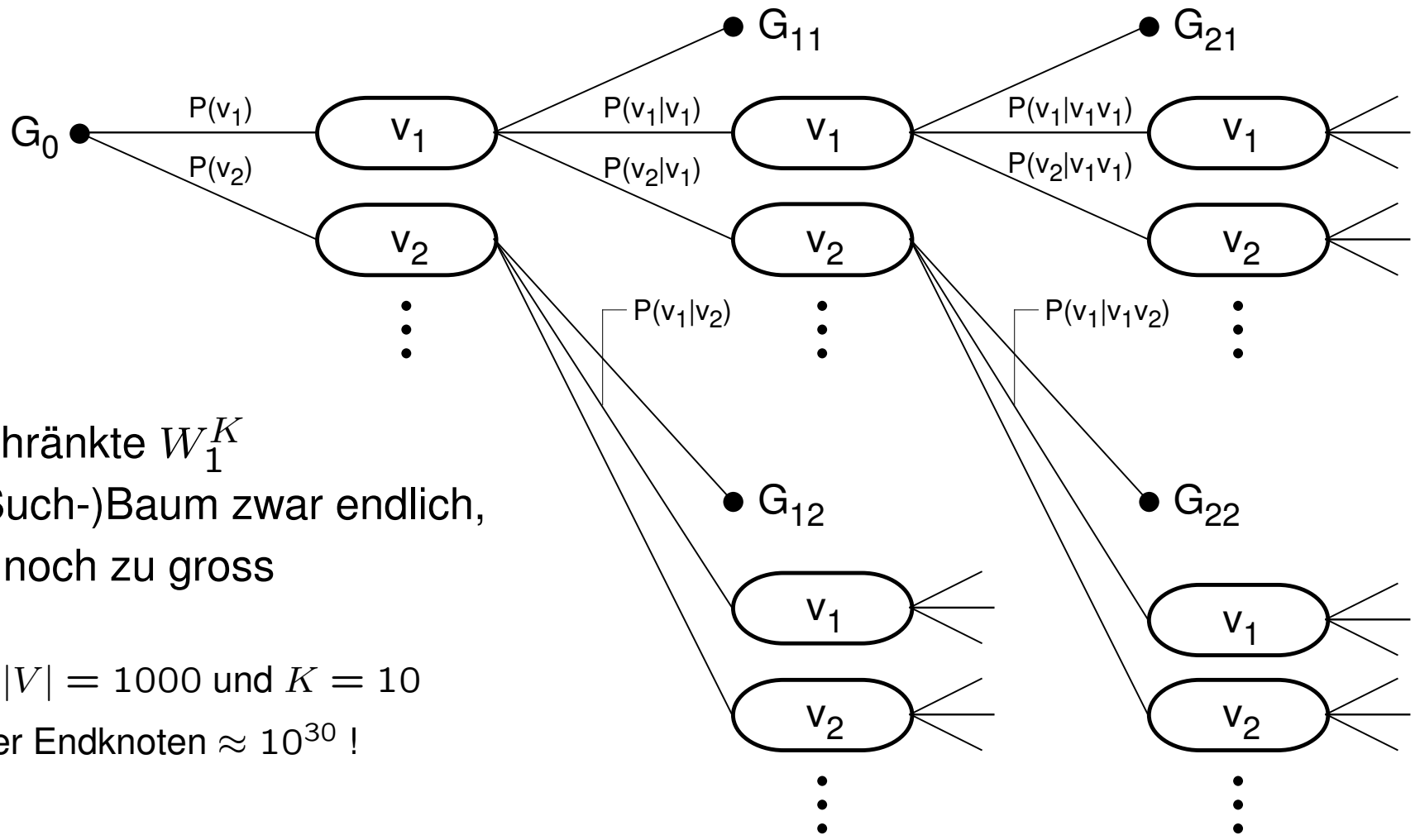
$\longrightarrow$ *allgemeines statistisches Sprachmodell*

Problem: Menge aller Wortfolgen V^* ist i.a. sehr gross.

Ansatz: Zerlegung in bedingte Wahrscheinlichkeiten:
$P(w_1 \dots w_K) = P(w_1) \cdot P(w_2|w_1) \cdots P(w_K|w_1 \dots w_{K-1})$

Anwendung: Dank kausaler Zerlegung Baumdarstellung möglich

Allgemeines statistisches Sprachmodell



Für beschränkte W_1^K
ist der (Such-)Baum zwar endlich,
aber dennoch zu gross

z.B. ist für $|V| = 1000$ und $K = 10$
die Zahl der Endknoten $\approx 10^{30}$!

N-Gram-Sprachmodell

Problem: Das allgemeine statistische Sprachmodell
$P(w_1 \dots w_K) = P(w_1) \cdot P(w_2|w_1) \cdots P(w_K|w_1 \dots w_{K-1})$
ist zu gross!

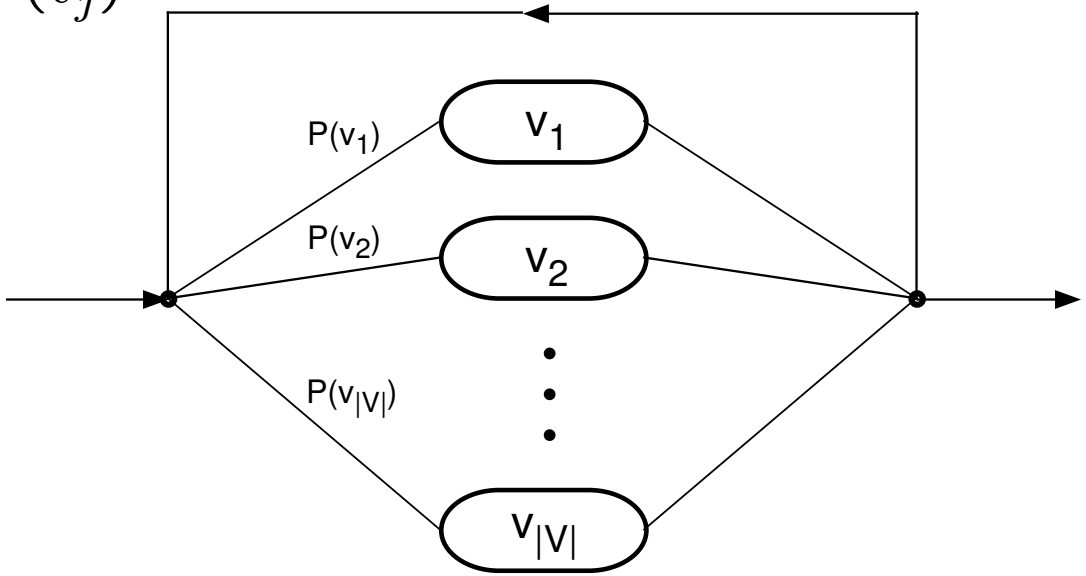
Beobachtung: Weit auseinander liegende Wörter sind i.a. nur schwach voneinander abhängig

Ansatz: Approximation der bedingten Wahrscheinlichkeiten, indem nur die letzten $N-1$ Vorgänger des aktuellen Wortes betrachtet werden:
$P(w_k|w_1 \dots w_{k-1}) \approx P(w_k|w_{k-N+1} \dots w_{k-1})$

Unigram-Sprachmodell: $N = 1$

Die Abhängigkeit des Wortes w_k von den vorangehenden wird vernachlässigt:
$P(w_k|w_1 \ldots w_{k-1}) \approx P(w_k)$

$\longrightarrow$ Unigram-Modell berücksichtigt nur die A-priori-Wahrscheinlichkeit der Wörter: $P(v_j)$

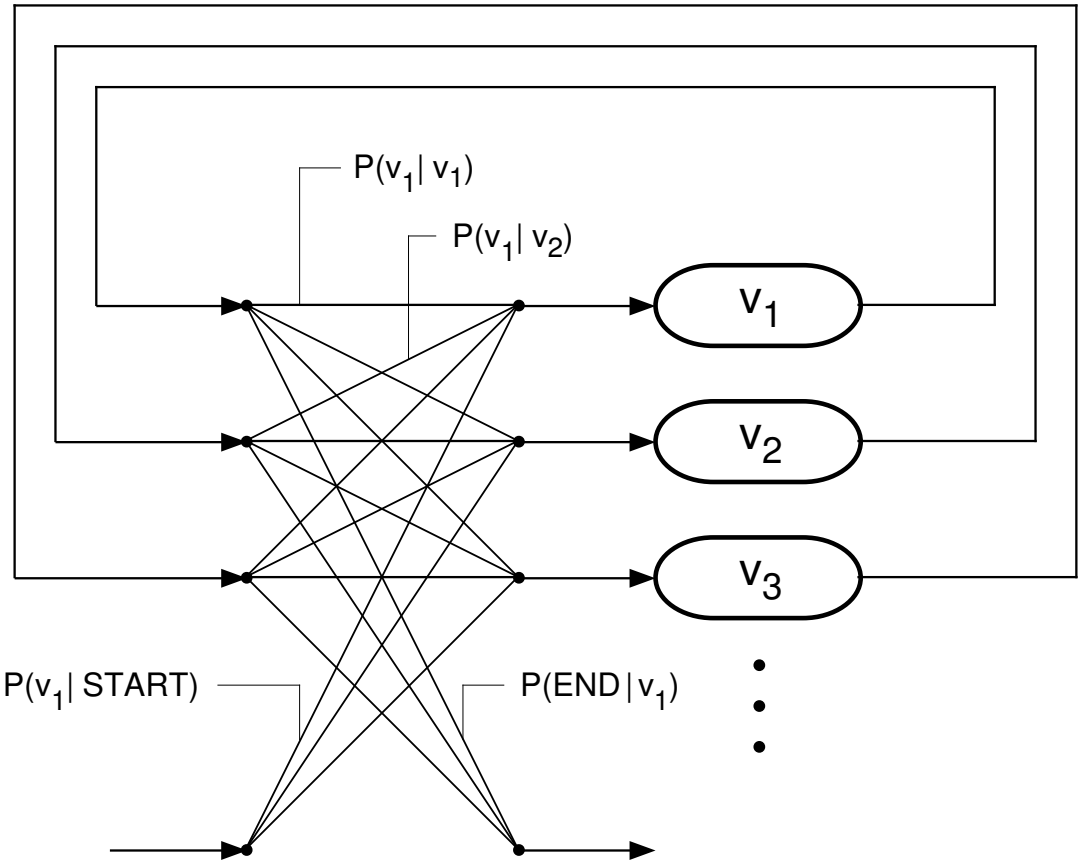


Bigram-Sprachmodell: $N = 2$

Es wird das Wort w_k als nur von seinem direkten Vorgänger abhängig betrachtet:

$$P(w_k|w_1 \ldots w_{k-1}) \approx P(w_k|w_{k-1})$$

$\longrightarrow$ Bigram-Modell
mit $(|V| + 1)^2 - 1$
Wahrscheinlichkeiten
$P(w_k = v_j \,|\, w_{k-1} = v_i)$



Trigram-Sprachmodell

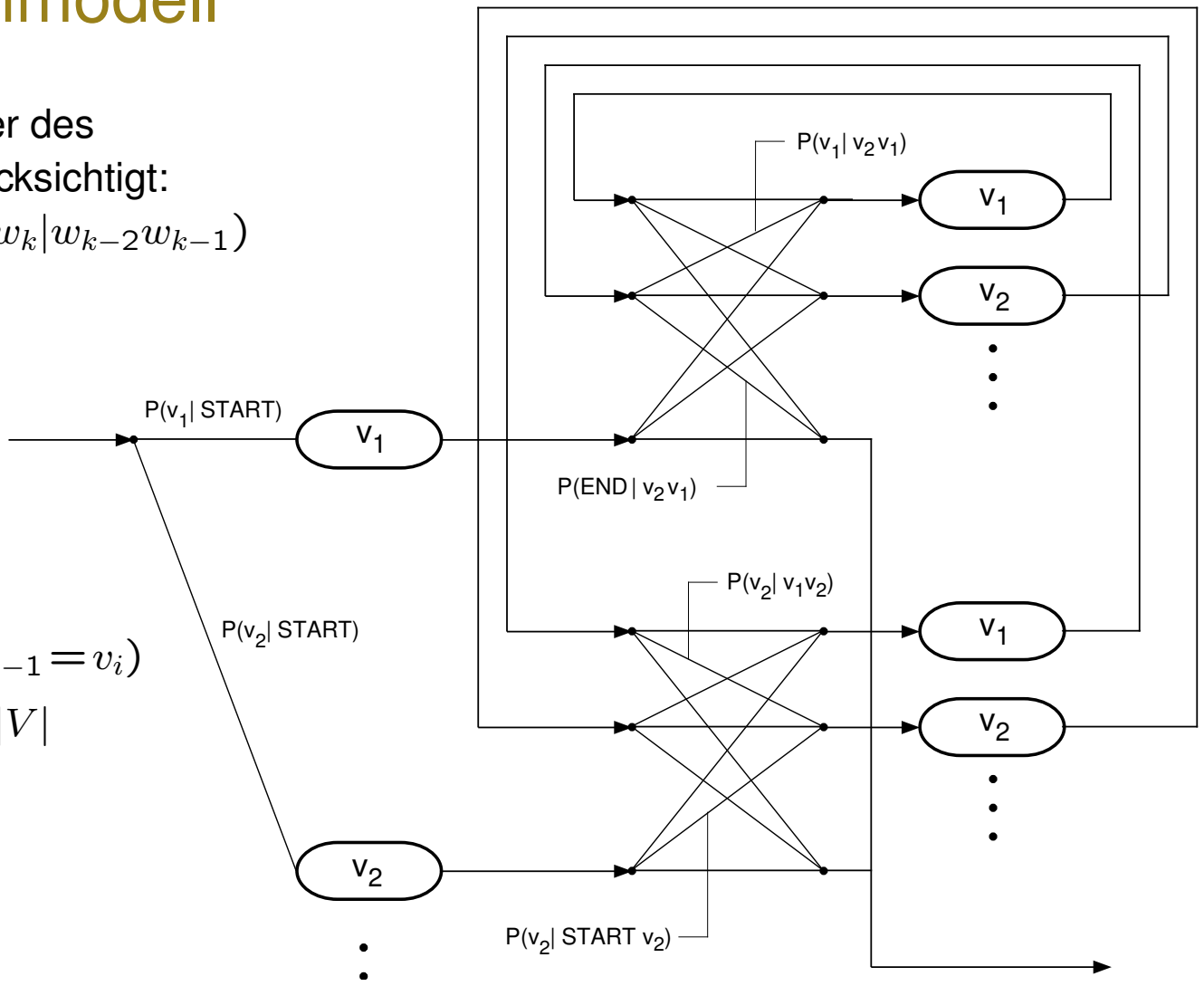
Es werden zwei Vorgänger des aktuellen Wortes w_k berücksichtigt:

$$P(w_k|w_1 \ldots w_{k-1}) \approx P(w_k|w_{k-2}w_{k-1})$$

Anzahl Modellparameter d.h. Wahrscheinlichkeiten

$$P(w_k = v_j \mid w_{k-2} = v_h, w_{k-1} = v_i)$$

$$\longrightarrow \quad |V| + (|V|+1)^2 \cdot |V|$$



Schätzen der N-Gram-Sprachmodell-Parameter

Methode: Auszählen, wie oft im Stichprobentext auf die Wortfolge $h^{(N-1)}$ das Wort w folgt:

$$\tilde{P}(w|h^{(N-1)}) = \mathsf{freq}(w|h^{(N-1)}) = \frac{\mathsf{cnt}(h^{(N-1)}w)}{\sum_{w'} \mathsf{cnt}(h^{(N-1)}w')}$$

Problem: nicht vorhandene Wortfolgen $h^{(N-1)}w$ (oder Wortfolgen $h^{(N-1)}$)

$\longrightarrow \tilde{P}(w|h^{(N-1)}) = 0$ (bzw. nicht definiert)

Konsequenz: Wortfolgen, die $h^{(N-1)}w$ enthalten, werden nicht erkannt!

Abhilfe: Glättung mit: $\tilde{P}(w|h^{(N-1)}) = \frac{\mathsf{cnt}(h^{(N-1)}w) + \alpha}{\sum_{w'}(\mathsf{cnt}(h^{(N-1)}w') + \alpha)}$

wobei: $0 < \alpha < 0.5$

Rückgriff auf einfachere Sprachmodelle

a) Lineare interpolation

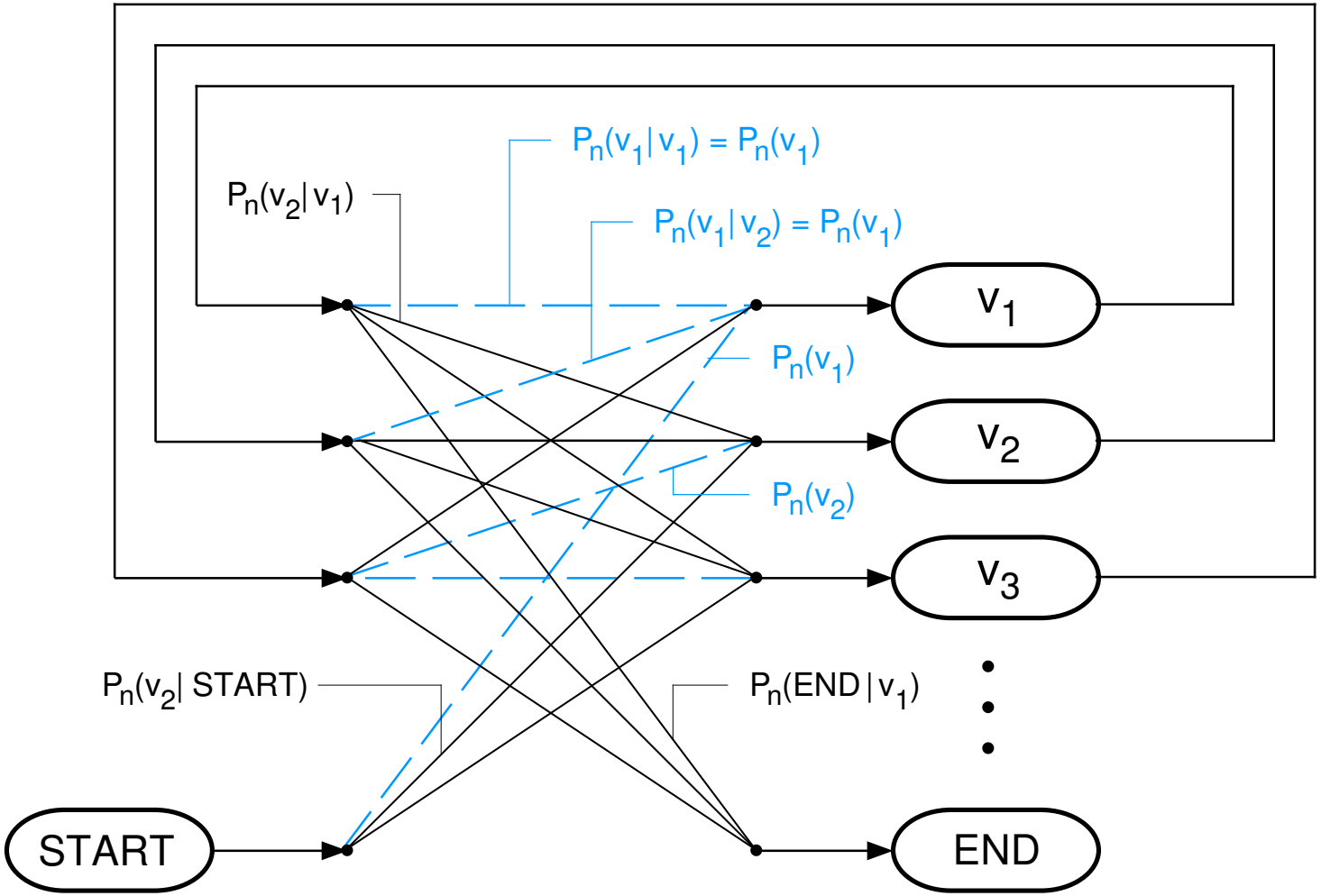
$$\tilde{P}_s(w|h^{(N-1)}) \;=\; \rho_1 \tilde{P}(w) + \rho_2 \tilde{P}(w|h^{(1)}) + \ldots + \rho_N \tilde{P}(w|h^{(N-1)})$$

(ρ_i aus einem andern Stichprobentext schätzen)

b) Backing-off

Für fehlende N-Gram-Wahrscheinlichkeiten wird auf (N-1)-Gram-Wahrscheinlichkeiten zurückgegriffen

Backing-off-Bigram-Sprachmodell

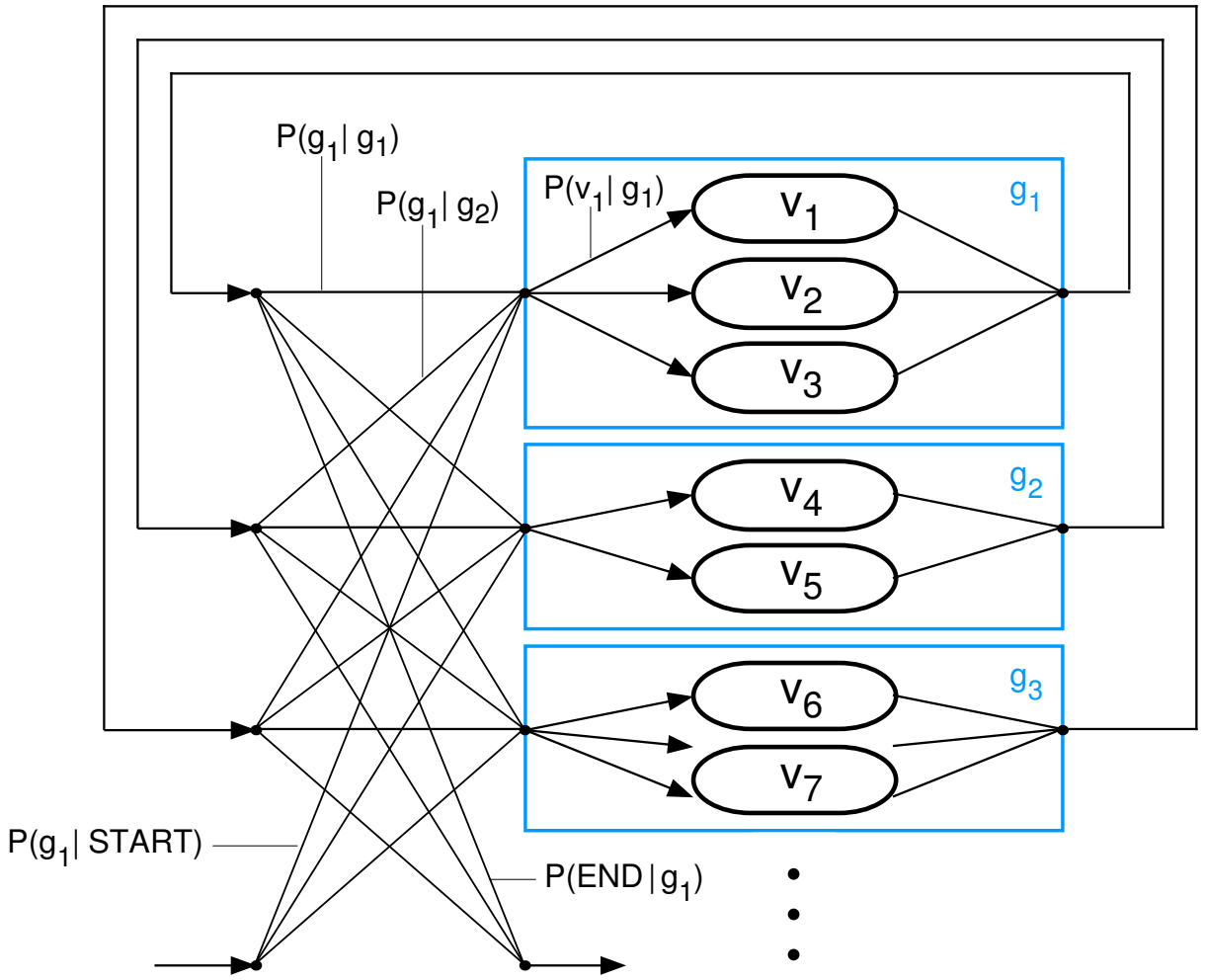


Kategorielle N-Gram-Sprachmodelle

Beispiel: kateg. Bigram

Kategorien-N-Gram mit
mit Unigram pro Kategorie

Vorteilhaft, da
Anzahl Kategorien $\ll |V|$



Bewertung von Sprachmodellen

Optimal ist das Sprachmodell eines Spracherkenners dann, wenn es alle A-priori-Kenntnisse über die zu erkennende Sprache enthält.

Frage: Wie lässt sich ein Sprachmodell bewerten?

Antwort:

- Grundsätzlich ist das Sprachmodell umso besser, je mehr es die Entscheidungsfreiheit des Erkenners einschränkt.
- Gemessen wird dies anhand der *Entropie* oder der *Perplexität*

Information

Definition: Die Information des Wortes $v_i \in V = \{v_1, v_2, \ldots, v_{|V|}\}$ ist:

$$I(v_i) = -\log_2 P(v_i)$$

Für eine Wortfolge $W = w_1 w_2 \ldots w_K$ gilt:

- Die Informationszunahme durch das k-te Wort, gegeben W_1^{k-1}, ist:

  $$I_k(v_i) = -\log_2 P(w_k = v_i \,|\, w_1 w_2 \ldots w_{k-1})$$

- Die Gesamtinformation der Wortfolge $w_1 w_2 \ldots w_K$:

$$I(W_1^K) = -\log_2 P(w_1 w_2 \ldots w_K) = -\log_2 \prod_{k=1}^{K} P(w_k \,|\, w_1 w_2 \ldots w_{k-1})$$

$$= -\sum_{k=1}^{K} \log_2 P(w_k \,|\, w_1 w_2 \ldots w_{k-1}) = \sum_{k=1}^{K} I_k(w_k)$$

Entropie

Definition: Mittlerer Informationszuwachs pro Wort:

$$H = \lim_{K\to\infty} \frac{1}{K} I(W_1^K) = -\lim_{K\to\infty} \frac{1}{K} \log_2 P(W_1^K)$$

Für eine Wortfolge W ist die bedingte Entropie des k-ten Wortes:

$$\begin{aligned} H(w_k|W_1^{k-1}) &= \sum_{i=1}^{|V|} P(v_i|W_1^{k-1})\, I(v_i|W_1^{k-1}) \\ &= -\sum_{i=1}^{|V|} P(v_i|W_1^{k-1}) \log_2 P(v_i|W_1^{k-1}) \end{aligned}$$

Perplexität

Definition: $\mathcal{Q} = 2^H$

(Verzweigungsrate)

>>>

Entropie eines N-Gram-Sprachmodells

Erwartungswert der bedingten Entropie des k-ten Wortes
(über alle Kombinbationen der $N-1$ vorangegangenen Wörter)

$$H_{mod} = \sum_{W_{k-N+1}^{k-1}} P(W_{k-N+1}^{k-1})\, H(w_k|W_{k-N+1}^{k-1})$$

$$= -\sum_{v_h,v_i,\ldots,v_k,v_l \in V} P(v_h v_i \ldots v_k v_l)\, \log_2\, P(v_l|v_h v_i \ldots v_k)$$

Entropie als Gütekriterium

Feststellung: In der Sprache L mit dem Vokabular V sind die Wörter nicht gleich häufig.

Folge 1: Für die Entropie der Sprache und des Unigram-Modells gilt:

$$H_L < \sum_{i=1}^{|V|} P(v_i) I(v_i) < -\log_2 \frac{1}{|V|} = \log_2 |V|$$

Folge 2: Das Sprachmodell ist umso besser, je mehr Abhängigkeiten es beschreibt.

Gütekriterium: $\frac{H_L}{H_{mod}} \leq 1$ (optimal = 1)

>>>

Entropie als Gütekriterium

Problem: Sprachmodell kann zu einschränkend sein, sodass $H_{mod} < H_L$
(z.B. weil gewisse Wortfolgen im Trainingstext fehlen)

Lösung: Güte des Sprachmodells mittels der Kreuzentropie aus einem K Wörter langen Testtext W_1^K ermitteln:

$$H(W_1^K) = -\frac{1}{K} \sum_{k=1}^{K} log_2 \; P(w_k | w_{k-N+1} \dots w_{k-1})$$

Merke: eine im Testtext einmal vorkommende Wortfolge hat die Häufigkeit $\frac{1}{K}$

Stärken statistischer Sprachmodelle

N-Gram sind leistungsfähig:

- sie werden mit in einer Anwendung wirklich gebrauchten Sätzen trainiert (Anwendungsrelevanz)
- sie beinhalten syntaktisches, semantisches und pragmatisches Wissen (in statistischer Form)

Trainierbarkeit:

N-Gram lernen die relevante Information selbständig aus den Trainingstexten

- kein linguistisches Wissen über (ev. fremde) Sprachen nötig
- Sprachunabhängigkeit

Effizient kombinierbar mit akustischen Modellen:

- N-Gram $\longrightarrow$ Übergangwahrscheinlichkeiten im Erkennungsnetzwerk
- effiziente Verarbeitung mit Viterbi-Algorithmus

Schwächen statistischer Sprachmodelle

Modellschwäche:

Aus praktischen Gründen ist nur ein kurzer Wortkontext handhabbar:

- Umfang des Stichprobentextes zu limitiert
- Erkennungsnetzwerk zu gross

Mangelnde linguistische Relevanz:

Die Unterscheidung zwischen sprachlich korrekten und falschen Sätzen ist ungenügend.

Überbewertung des Sprachmodells:

Das Sprachmodell kann das akustische Modell permanent überstimmen (Beispiel: “Möller” und “Müller”)

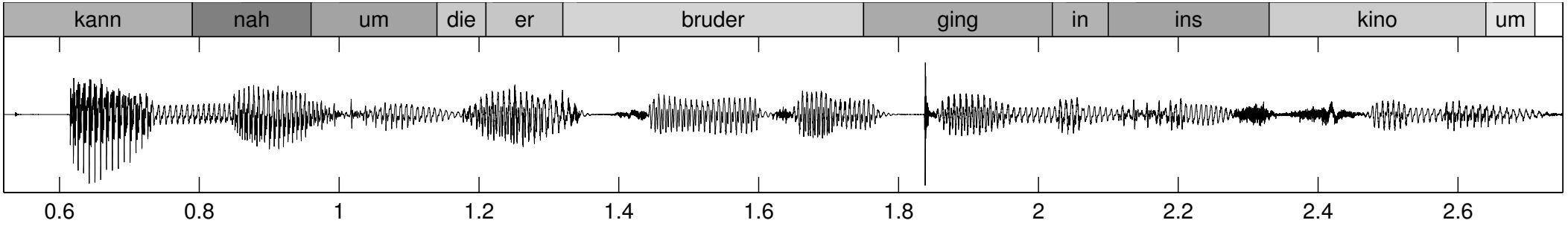
Thema der nächsten Lektion:

Wissenbasierte Sprachmodelle

Zur Übersicht der Vorlesung *Sprachverarbeitung II* >>>

Beispiel

Viterbi-Erkennung einer Wortfolge mit einem *Word-Loop*-Erkennungsnetzwerk
(Wortfolge aus der optimalen Zustandssequenz)



<<<

Rechenbeispiel

Gegeben: Sprache $L_a = \{w_a | w_a \in \{a, b\}\}$, wobei $P(a) = P(b)$

Entropie eines Wortes w: $H(w) = -\sum_{i=1}^{|V|} P(v_i) \log_2 P(v_i)$

$$H(w_a) = -\tfrac{1}{2} \log_2 (\tfrac{1}{2}) - \tfrac{1}{2} \log_2 (\tfrac{1}{2}) = 1$$

Perplexität von w: $\mathcal{Q}(w) = 2^{H(w)}$

$$\mathcal{Q}(w_a) = 2^1 = 2$$

Rechenbeispiel

Gegeben: Sprache $L_a = \{w_a | w_a \in \{a, b\}\}$, wobei $P(a) = P(b)$

Sprache $L_b = \{w_b | w_b \in \{a, b\}\}$, wobei $P(a) = 2P(b)$

Entropie eines Wortes w: $H(w) = -\sum_{i=1}^{|V|} P(v_i) \log_2 P(v_i)$

$$H(w_a) = -\tfrac{1}{2} \log_2 (\tfrac{1}{2}) - \tfrac{1}{2} \log_2 (\tfrac{1}{2}) = 1$$

$$H(w_b) = -\tfrac{2}{3} \log_2 (\tfrac{2}{3}) - \tfrac{1}{3} \log_2 (\tfrac{1}{3}) = 0.918$$

Perplexität von w: $\mathcal{Q}(w) = 2^{H(w)}$

$$\mathcal{Q}(w_a) = 2^1 = 2$$

$$\mathcal{Q}(w_b) = 2^{0.918} = 1.89$$

<<<

Rechenbeispiel

Gegeben: Sprache $L = \{w \in \{a,b\}^*\}$,
wobei $P(a) = 2P(b)$, $P(a|a) = 3P(b|a)$ und $P(bb) = 0.1$

Gesucht: geeignete statistische Grammatik (N-Gram)

Lösung: N-Gram mit $H_{mod} \approx H(L)$ wobei

$$H_{mod} = -\sum_{w \in V} P(w_{k-N+1} \ldots w_k) \log_2 P(w_k | w_{k-N+1} \ldots w_{k-1})$$

N-Gram-Entropie:

$H_0 = -\log_2 \frac{1}{2} = 1$

$H_1 = -P(a) \log_2 P(a) - P(b) \log_2 P(b) = 0.918$

$H_2 = -P(aa) \log_2 P(a|a) - P(ab) \log_2 P(b|a)$
$\quad -P(ba) \log_2 P(a|b) - P(bb) \log_2 P(b|b) = 0.835$

Perplexität:

$\mathcal{Q}_0 = 2^{H_0} = 2$

$\mathcal{Q}_1 = 2^{H_1} = 1.89$

$\mathcal{Q}_2 = 2^{H_2} = 1.78$

<<<

Rechenbeispiel (Fortsetzung)

Aus den Vorgaben über L und mit

$$P(aa) = P(a|a)\,P(a) \qquad P(a) + P(b) = 1$$
$$P(ab) = P(b|a)\,P(a) \qquad P(a|a) + P(b|a) = 1$$
$$P(ba) = P(a|b)\,P(b) \qquad P(a|b) + P(b|b) = 1$$
$$P(bb) = P(b|b)\,P(b) \qquad P(aa) + P(ab) + P(ba) + P(bb) = 1$$

können alle nötigen Wahrscheinlichkeiten berechnet werden:

$$P(a) = 1 - P(b) = 1 - \tfrac{1}{2}P(a) = \tfrac{2}{3} \qquad P(b) = \tfrac{1}{3}$$
$$P(a|a) = 1 - P(b|a) = 1 - \tfrac{1}{3}P(a|a) = \tfrac{3}{4} \qquad P(b|a) = \tfrac{1}{4}$$
$$P(aa) = P(a|a)\,P(a) = \tfrac{3}{4}\,\tfrac{2}{3} = \tfrac{1}{2} \qquad P(bb) = \tfrac{1}{10}$$
$$P(b|b) = P(bb)/P(b) = \tfrac{1}{10}\,\tfrac{3}{1} = \tfrac{3}{10} \qquad P(a|b) = 1 - P(b|b) = \tfrac{7}{10}$$
$$P(ab) = P(b|a)\,P(a) = \tfrac{1}{4}\,\tfrac{2}{3} = \tfrac{1}{6} \qquad P(ba) = P(a|b)\,P(b) = \tfrac{7}{10}\,\tfrac{1}{3} = \tfrac{7}{30}$$

(Wortanfänge und -enden vernachlässigt)

<<<