7 Sprachmodellierung

7.1 Zum Begriff Sprachmodellierung in der Spracherkennung

Sprachmodell (language model)

- Sammlung von *A-priori-Kenntnissen* über die Sprache
- In diesem Zusammenhang wird Sprache im abstrakten Sinne verstanden, nämlich als eine Menge von Wortfolgen ohne bestimmte akustische Realisierung.
- Diese Kenntnisse stehen im Voraus zur Verfügung, also bevor eine zu erkennende lautliche Äußerung vorliegt.
- Wissen über die Sprache allgemein
- Wissen über die Kommunikationssituation
  - welche Wörter oder Wortfolgen werden in der gegebenen Situation gebraucht werden und wie häufig
- Das Sprachmodell eines Spracherkenners besteht meistens aus mehreren Teilmodellen.
  - Vokabular
  - Häufigkeit der Wörter
  - Satzgrammatik

Modellformen

- **Statistische Sprachmodelle:**
  - Ermittelt man die Erfahrungswerte durch Messen oder Auszählen von Ereignissen in einer Sprachdatensammlung, z.B. die Häufigkeit oder die Auftretenswahrscheinlichkeit eines Wortes, dann sprechen wir von einem statistischen Sprachmodell.
  - Ein statistisches Sprachmodell beschreibt die Sprache so, als würde sie durch einen Zufallsprozess erzeugt.

- **Wissensbasierte Sprachmodelle:**
  - Wird linguistisches Expertenwissen in einem Sprachmodell angewendet, beispielsweise grammatikalisches Wissen, dann handelt es sich um ein wissensbasiertes Sprachmodell.
  - Grundlage dieses Wissens ist weniger die zahlenmäßige Erfassung von Beobachtungen der sprachlichen Oberfläche, sondern eher Einsichten über die hinter diesen Beobachtungen stehenden Zusammenhänge und Gesetzmäßigkeiten der Sprache.

7.2 Statistische Sprachmodellierung

MAP – Regel (Maximum-a-posteriori-Regel

$$\hat{W} = \underset{W \in V^*}{\operatorname{argmax}} P(W \mid \mathbf{X})$$

optimale Wortfolge

$\hat{W} = w_1 w_2 \ldots w_K \quad w_i \in V$

Wortfolge

Merkmalssequenz

$\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_T$

$$P(W \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid W) P(W)}{P(\mathbf{X})}$$

$$\hat{W} = \underset{W \in V^*}{\operatorname{argmax}} P(\mathbf{X} \mid W) \cdot P(W)$$

akustisches Modell
z.B.: HMM

Sprachmodell

Sprachmodell

Wahrscheinlichkeitsverteilung: $P(W)$

Es ist von der Merkmalssequenz **X** unabhängig und beinhaltet deshalb ausschließlich die oben erwähnten A-priori-Kenntnisse.

alle Wortfolgen sind grundsätzlich möglich, aber manche sind wahrscheinlicher als andere

Häufigkeiten einzelner Wörter

Rang	Wort	Häufigkeit in %
1	*die*	3.288
2	*der*	3.208
3	*und*	2.885
4	*in*	1.499
5	*das*	1.283
6	*zu*	1.135
7	*ist*	1.049
8	*sie*	1.049
9	*den*	1.019
10	*nicht*	0.9116
11	*von*	0.9016
12	*ich*	0.8991
13	*es*	0.8141
14	*wir*	0.777
15	*mit*	0.7739

7.2.1 Sprachmodellierung bei der Einzelworterkennung

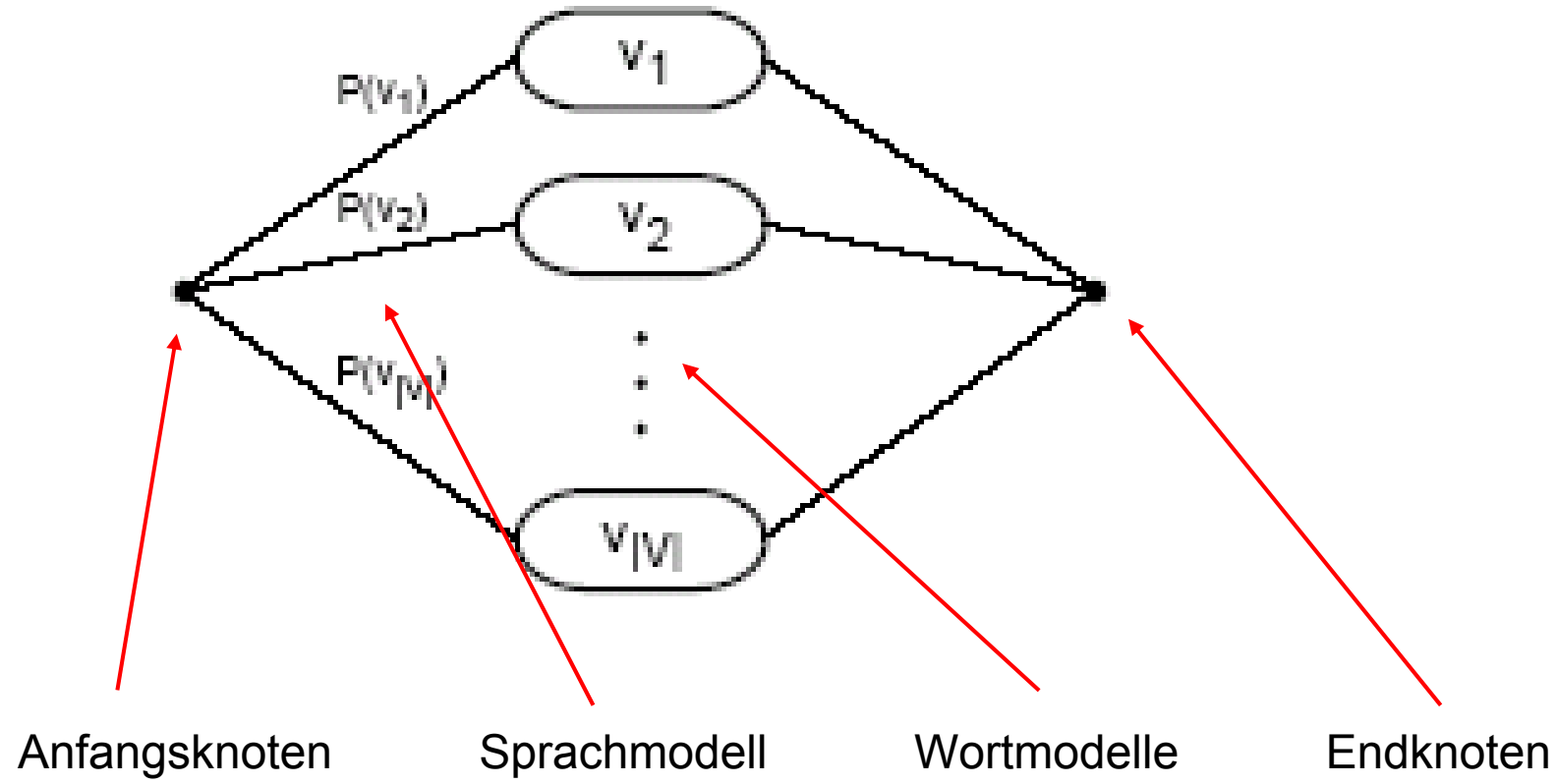
Sprachmodell

Wörter des Erkennervokabulars $V = \{v_1, v_2, \ldots, v_{|V|}\}$

Sprachmodell: $P(v_1)$ $P(v_2)$ $P(v_{|V|})$

A-priori-Wahrscheinlichkeit für jedes Wort

Netzwerk eines Spracherkenners für einzeln gesprochene Wörter



Beispiel – Fahrkartenautomat

Wenn die akustischen Modelle dieses Systems dazu neigen, die Ortsnamen “Bern” und “Berg” zu verwechseln, so führt das ohne Sprachmodell zu sehr vielen Erkennungsfehlern, da wesentlich mehr Menschen eine Fahrkarte in die Hauptstadt der Schweiz kaufen möchten als in eine der beiden kleinen Schweizer Gemeinden letzteren Namens.

Mit A-priori-Wahrscheinlichkeiten für die einzelnen Ziele lässt sich die Anzahl der Erkennungsfehler minimieren.
Die entsprechenden Wahrscheinlichkeiten könnte man leicht aus den Statistiken schätzen, die von der früheren Generation von Fahrkartenautomaten gesammelt wurden:

Um das Sprachmodell noch weiter zu verfeinern, könnte man die A-priori-Wahrscheinlichkeiten vom Standort des Automaten abhängig machen. Das würde dann vermutlich dazu führen, dass in den Nachbargemeinden von “Berg” tendenziell eher “Berg” erkannt würde als “Bern”.

Beispiel – Dialog

“Womit kann ich Ihnen dienen?”
“Fahrkarte”
“Nennen Sie bitte Ihren Zielort!”
“Bern”
“Hin- und Rückfahrt?”
“Ja”
...

Das Sprachmodell wird hier selbstverständlich genauer, wenn je nach Dialogschritt unterschiedliche Wortstatistiken angewendet werden: Nach der Frage “Hin- und Rückfahrt?” sind die Antworten “ja”, “nein”, “klar” usw. viel wahrscheinlicher als etwa “Fahrkarte”.

7.2.2 Sprachmodellierung für Wortfolgen

Sprachmodell

Wortfolge: $W_1^K = w_1 w_2 \ldots w_K$

$$w_k \in V = \{v_1, \ldots v_{|V|}\}$$

Sprachmodell: $P(W_1^K)$

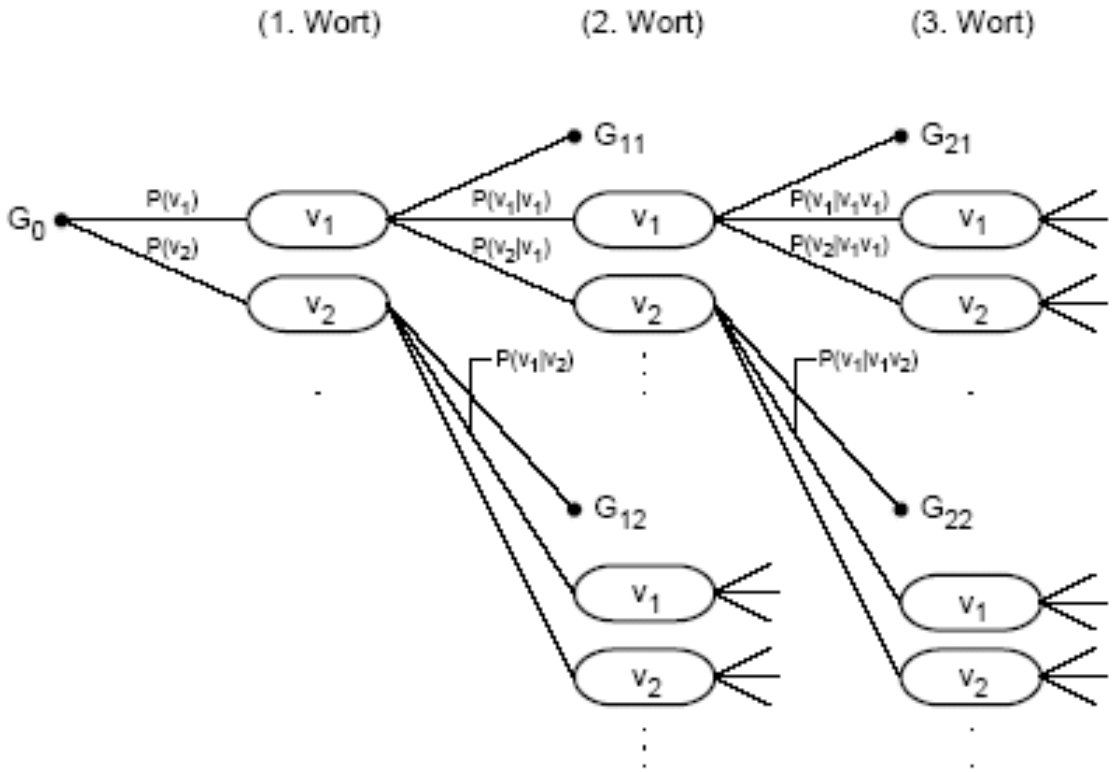
$$P(W_1^K) = P(w_1) \cdot P(w_2 \mid w_1) \cdots P(w_K \mid w_1 \ldots w_{K-1})$$

$$= \prod_{k=1}^{K} P(w_k \mid W_1^{k-1})$$

7.2.3 Das allgemeine statistische Sprachmodell

- Die allgemeine Formulierung des statistischen Sprachmodells als Wahrscheinlichkeit P(W) für jede endliche Wortfolge W Є V ＊ übt auf die Art der Sprache keinerlei Einschränkungen aus.
- Eine solche Beschreibung der Sprache umfasst alle Sprachebenen (die lexikalische, die syntaktische, die semantische und die pragmatische Ebene), denn sie kann z.B. syntaktisch korrekten Wortfolgen gegenüber inkorrekten durch eine höhere Wahrscheinlichkeit den Vorzug geben, aber auch genauso semantisch und pragmatisch sinnvolle Wortfolgen bevorzugen.

Allgemeines Sprachmodell – Suchbaum



Die Wahrscheinlichkeit der Wortfolge W_1^K ergibt sich aus dem Produkt der bedingten Wahrscheinlichkeiten der am Pfad beteiligten Kanten.

$$m = 1, \ldots, |V|^K$$

$$W_1^K = w_1 w_2 \ldots w_K \longrightarrow G_0 \rightarrow G_{Km} \longrightarrow \text{(Endknoten)}$$

Weg über $w_j = v_{i_j}$

Einschätzung

- Wenn wir aber bedenken, dass bereits bei einer bescheidenen Vokabulargröße von $|V|=1000$ und einer maximalen Länge der Wortfolgen von 10 die Zahl der Endknoten $=1000^{10}=10^{30}$ beträgt, dann ist schon das Speichern und erst recht das Schätzen der bedingten Wahrscheinlichkeiten unmöglich.
- Das allgemeine statistische Sprachmodell entpuppt sich somit als rein theoretischer Ansatz zur Sprachmodellierung, dem in der Praxis überhaupt keine Bedeutung zukommt.

7.2.4 N-Gram Sprachmodelle

Approximation

- man kann vereinfachend annehmen, dass die Wahrscheinlichkeit eines Wortes nicht von allen vorangegangenen, sondern näherungsweise nur von den letzten *N*−1 Wörtern abhängt

$$P(w_k \mid w_1 \ldots w_{k-1}) \approx P(w_k \mid w_{k-N+1} \ldots w_{k-1})$$

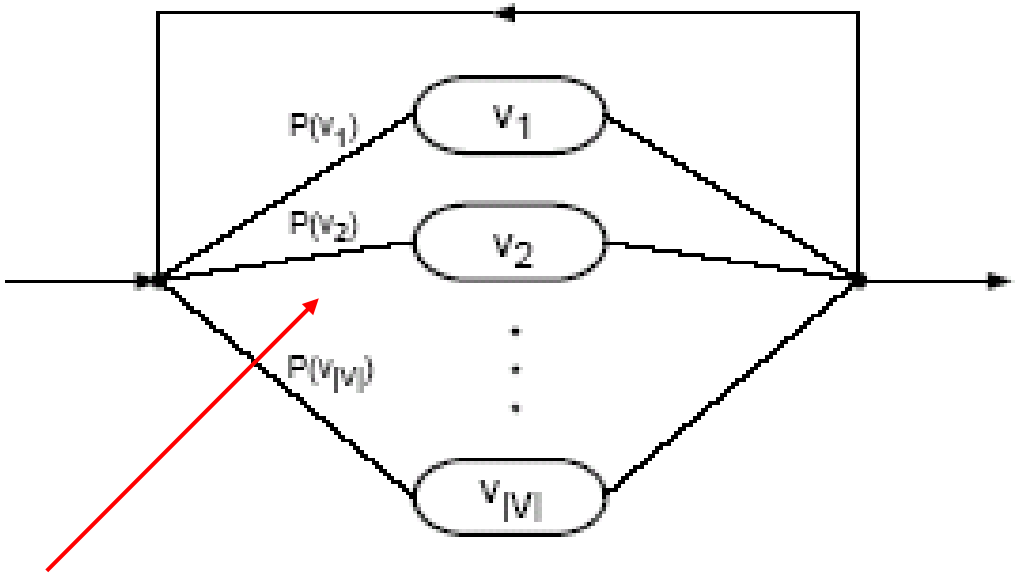
Wir betrachten die Fälle N=1 (Unigram), N=2 (Bigram) und N=3 (Trigram).

Unigram – Sprachmodell

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

In einem Unigram-Sprachmodell für Wortfolgen *W* aus dem Vokabular *V* steckt somit dieselbe Information wie in einem Sprachmodell zur Erkennung von einzeln gesprochenen Wörtern aus dem Vokabular *V.*

Unigram-Netzwerk eines Spracherkenners für beliebig lange Wortfolgen



Das Bigram – Sprachmodell

- In der Praxis wird häufig das Bigram-Sprachmodell verwendet.
- Bei diesem Sprachmodell wird angenommen, dass das aktuelle Wort w_k einer Wortfolge näherungsweise nur vom Vorgängerwort abhängt:

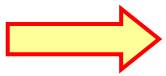
$$P(w_k \mid w_1 \ldots w_{k-1}) \approx P(w_k \mid w_{k-1})$$

Wortpaar Häufigkeiten

Rang	Wortpaar	Häufigkeit in %
1	*in der*	0.2849
2	*bei der*	0.2398
3	*für die*	0.1945
4	*in den*	0.1419
5	*und der*	0.1377
6	*und die*	0.1282
7	*das ist*	0.1233
8	*auf die*	0.1022
9	*von der*	0.1015
10	*mit dem*	0.0970
11	*mit der*	0.0867
12	*in die*	0.0818
13	*dass die*	0.0746
14	*ist die*	0.0725
15	*es ist*	0.0721

Hilfswörter

- Typischerweise kommt ein bestimmtes Wort nicht gleich häufig am Anfang und am Ende der Wortfolge vor.
- Falls eine Wortfolge einem Satz entspricht, denke man beispielsweise an die Präposition “im”, die oft am Satzanfang steht, während sie am Satzende praktisch ausgeschlossen ist.



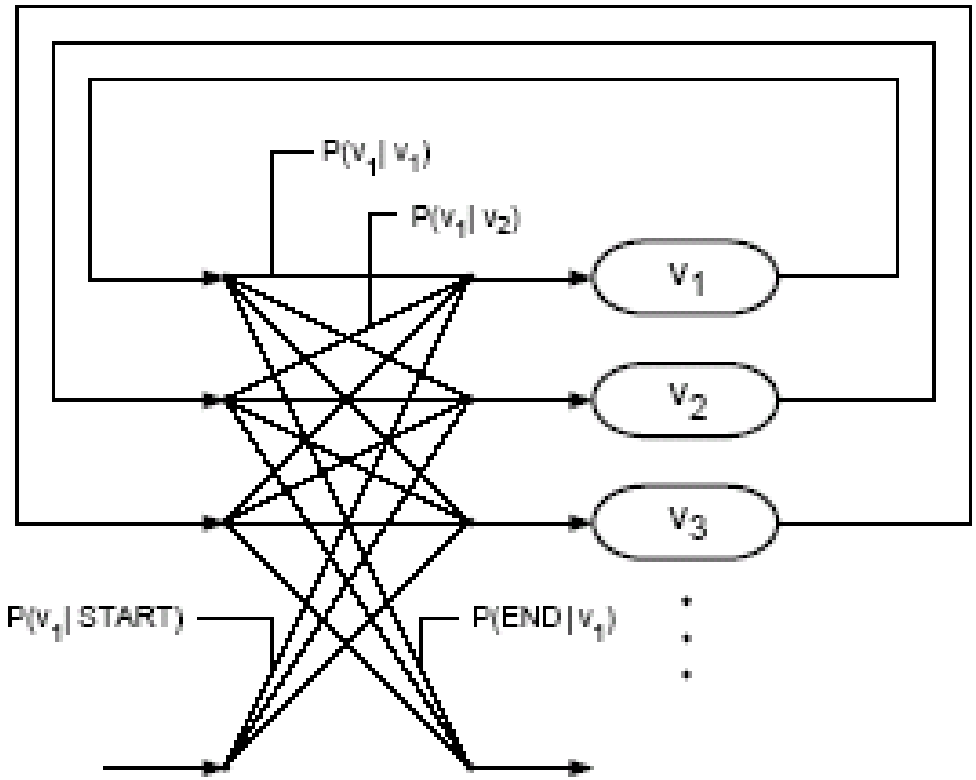
Hilfswörter

START

END

$$P(v_j \mid \text{START}) \qquad P(\text{END} \mid v_j) \qquad v_j \in V$$

Bigram-Netzwerk eines Spracherkenners für Wortfolgen

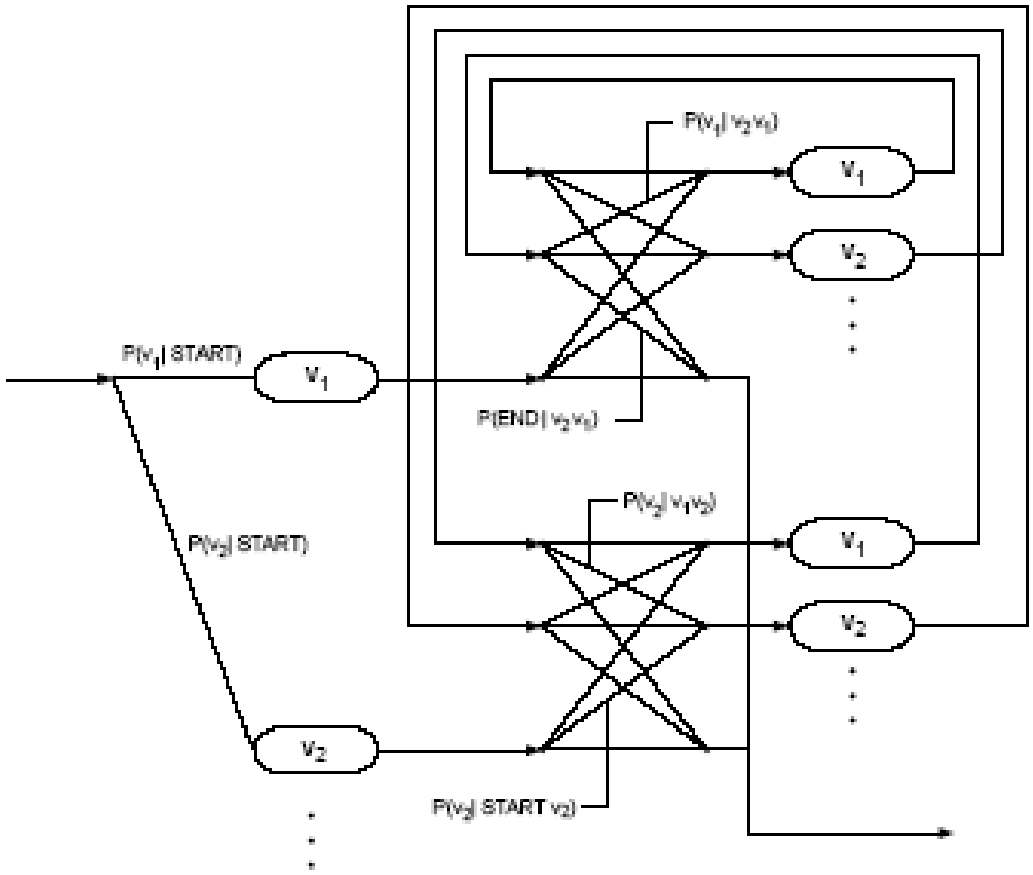


Das Trigram – Sprachmodell

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

Der Trigram-Ansatz stößt mit zunehmender Vokabulargröße schnell an praktische Grenzen, weil einerseits das Erkennungsnetzwerk unhandhabbar groß wird und andererseits die Trigram Wahrscheinlichkeiten schwierig zu schätzen sind. Viele Worttripel sind sehr selten.

Trigram-Netzwerk eines Spracherkenners für Wortfolgen



7.2.5 Schätzen der Parameter von N-Gram Sprachmodellen

Schätzung der bedingten Wahrscheinlichkeiten

- Die bedingten Wahrscheinlichkeiten eines N-Gram-Sprachmodells werden geschätzt, indem die relativen Häufigkeiten der betreffenden Wortfolgen aus einem großen Stichprobentext ermittelt werden.
- Für N=3 ist also die Wahrscheinlichkeit $P(v_k|v_iv_j)$ zu schätzen, indem gezählt wird, wie oft auf das Wortpaar v_iv_j das Wort v_k folgt.

$$P(v_k \mid v_i v_j) \approx \frac{anz(v_i v_j v_k)}{anz(v_i v_j)}$$

Probleme

- Das Problem dieser Schätzung ist, dass für im Stichprobentext nicht vorhandene Worttripel die bedingte Wahrscheinlichkeit null wird.
- Jede Wortfolge *W*, welche so eine Teilfolge enthält kann somit vom Spracherkenner nicht erkannt werden.
- Auch das Wortpaar $v_i v_j$ kann im Stichprobentext fehlen. Bedingte Wahrscheinlichkeit kann nicht geschätzt werden.
- Da dieses Problem beim praktischen Einsatz eines Spracherkenners überaus wichtig ist, werden im Folgenden verschiedene Methoden gezeigt, um mit diesem Problem umzugehen.
- Keine dieser Methoden löst das Problem grundsätzlich.

Glättung

- Dem Glätten (*smoothing*) liegt die Idee zu Grunde, dass gesehene Ereignisse etwas von ihrer Wahrscheinlichkeitsmasse an ungesehene abtreten, deren Wahrscheinlichkeiten damit nicht mehr null sind.
- Glätten der relativen Häufigkeiten

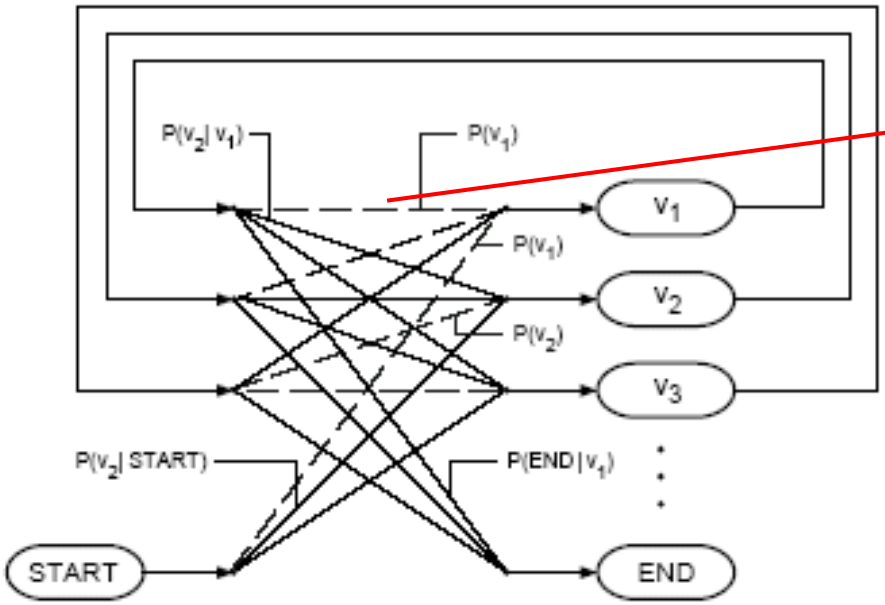
$$P(v_k \mid v_i v_j) \approx \frac{anz(v_i v_j v_k) + \alpha}{\sum_l (anz(v_i v_j v_l) + \alpha)}$$

kleine, positive Größe

Die Wahl ist jedoch kritisch und muss für jeden konkreten Anwendungsfall eingestellt werden.

Backing-off

- für fehlende N-Gram-Wahrscheinlichkeiten wird auf die entsprechenden N-1-Gram-Wahrscheinlichkeiten zurückgegriffen



Für Bigram-Wahrscheinlichkeiten, die null sind, sind die entsprechenden Verbindungen durch die Unigram- Wahrscheinlichkeiten ersetzt worden.

7.2.6 Kategorielle N-Gram-Sprachmodelle

Kategorielle N-Gram-Sprachmodelle

Wortkategorien $g_i \in G = \{g_1, \ldots, g_n\}$

$$P(w_k \mid w_{k-N+1} \ldots w_{k-1}) \approx \sum_{g_i \in G} P(w_k \mid g_i) P(g_i \mid w_{k-N+1} \ldots w_{k-1})$$

$$P(w_k \mid w_{k-N+1} \ldots w_{k-1}) \approx P(w_k \mid cat(w_{k-N+1}) \ldots cat(w_{k-1}))$$

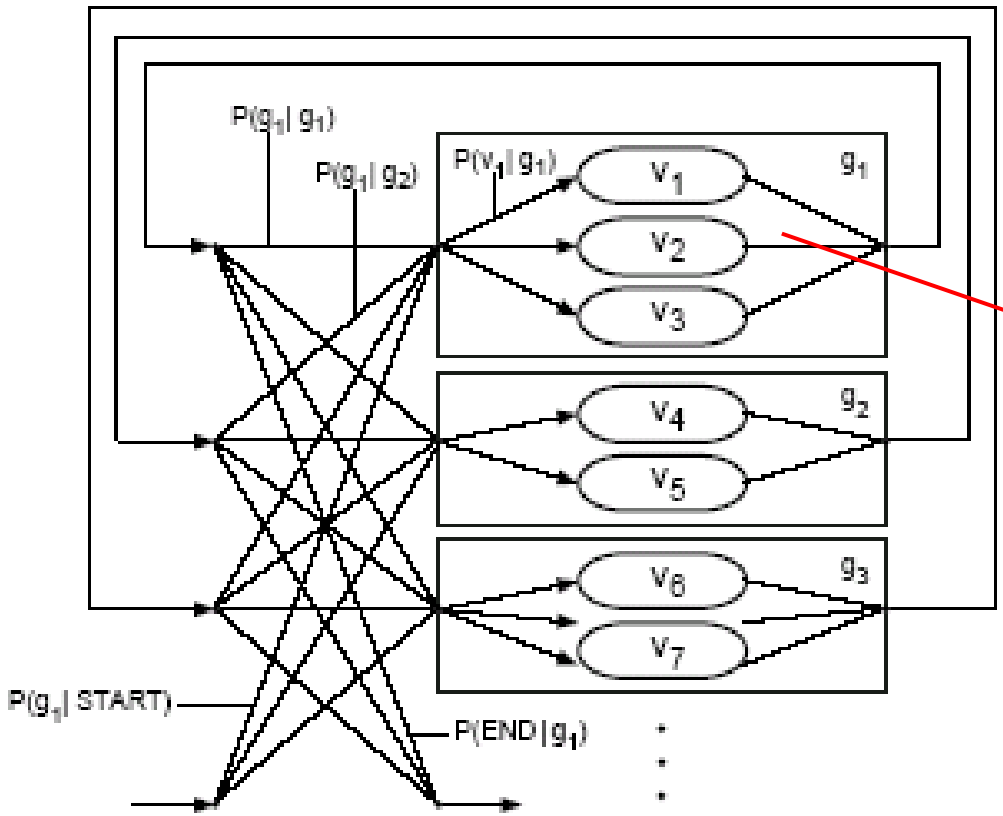
$$\approx \sum_{g_i \in G} P(w_k \mid g_i) P(g_i \mid cat(w_{k-N+1}) \ldots cat(w_{k-1}))$$

$$cat(w_i) \in G$$

Wortkategorien

- Prinzipiell können hier die grammatikalischen Wortkategorien (Nomen, Adjektiv, Artikel etc.) eingesetzt werden.
- Dazu müsste für alle Wörter eines Stichprobentextes die grammatikalische Wortkategorie bestimmt werden (um die Wahrscheinlichkeiten zu bestimmen). Abgesehen davon, dass dieses Vorgehen schwierig ist, führt es zu Kategorien sehr unterschiedlicher Größe.
- Es werden deshalb meistens künstliche Kategorien verwendet.
- Das heißt, es werden nicht nur obige Wahrscheinlichkeiten aus dem Stichprobentext geschätzt, sondern gleichzeitig auch die Wortkategorien bestimmt.

Kategorielles Bigram-Netzwerk eines Spracherkenners für Wortfolgen



Unigram Netzwerke

7.2.7 Bewertung von Sprachmodellen

Einführung

- Wie gut ist ein Sprachmodell bzw. um wieviel leichter wird die Erkennungsaufgabe durch das Sprachmodell?
- Mit Hilfe des Sprachmodells lassen sich aus der Vorgeschichte $w_1 w_2 . . . w_{k-1}$ Vorhersagen über das Wort w_k machen.
- Dabei gilt: je weniger sicher ein Wort aus seiner Vorgeschichte vorhersagbar ist,
  - desto mehr trägt das Wort zur Vermehrung der Information bei,
  - desto mehr Wörter kommen dort in Frage (große “Perplexität”),
  - desto größer ist dort die “Unvorhersagbarkeit” (große “Entropie”),
  - desto schwieriger ist dort die Erkennungsaufgabe.

Information

$$I(v_i) = -\log_2 P(v_i) \qquad V = \{v_1, v_2, \ldots, v_{|V|}\}$$

$$W_1^K = w_1 w_2 \ldots w_K \quad w_j = v_i \in V$$

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

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

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

Die Information der Wortfolge gleich der Summe der Information aller Wörter

Entropie

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

$$H_{\text{mod}} = -\sum_{v_h, v_i, \ldots, v_k, v_l \in V} P(v_h v_i \ldots v_k v_l) \log_2 P(v_l \mid v_h v_i \ldots v_k)$$

Die Entropie eines Sprachmodells ist umso kleiner, je sicherer sich das Modell (im Mittel) ist, welches Wort jeweils als nächstes kommt.

Je mehr Abhängigkeiten zwischen den einzelnen Wörtern das Modell berücksichtigt, desto geringer ist im Allgemeinen seine Entropie.

Kreuzentropie:

$$H(W_1^K) = -\frac{1}{K} \sum_{k=1}^{K} \log_2 P(w_k \mid w_1 w_2 \ldots w_{k-1})$$

Perplexität

$$Q = 2^{H}$$

Muss ein Spracherkenner beispielsweise Wortfolgen aus einem Vokabular der Größe |*V* | erkennen und es wird kein Sprachmodell eingesetzt, dann ist die Wahrscheinlichkeit jedes Wortes gleich groß, nämlich:

$$P(v_i) = \frac{1}{|V|} \qquad H_0 = H(W_1^K) = -\frac{1}{K}\log_2 P(W_1^K) = -\log_2 P(w_k) = \log_2 |V|$$

$$Q_0 = 2^{H_0} = |V|$$

$$Q_1 = 2^{H_{\text{mod}}}$$

zeigt um wieviel die mittlere Wortverzweigungsrate durch das Sprachmodell reduziert und damit die Spracherkennungsaufgabe vereinfacht wird.

7.2.8 Stärken und Schwächen der statistischen Modellierung

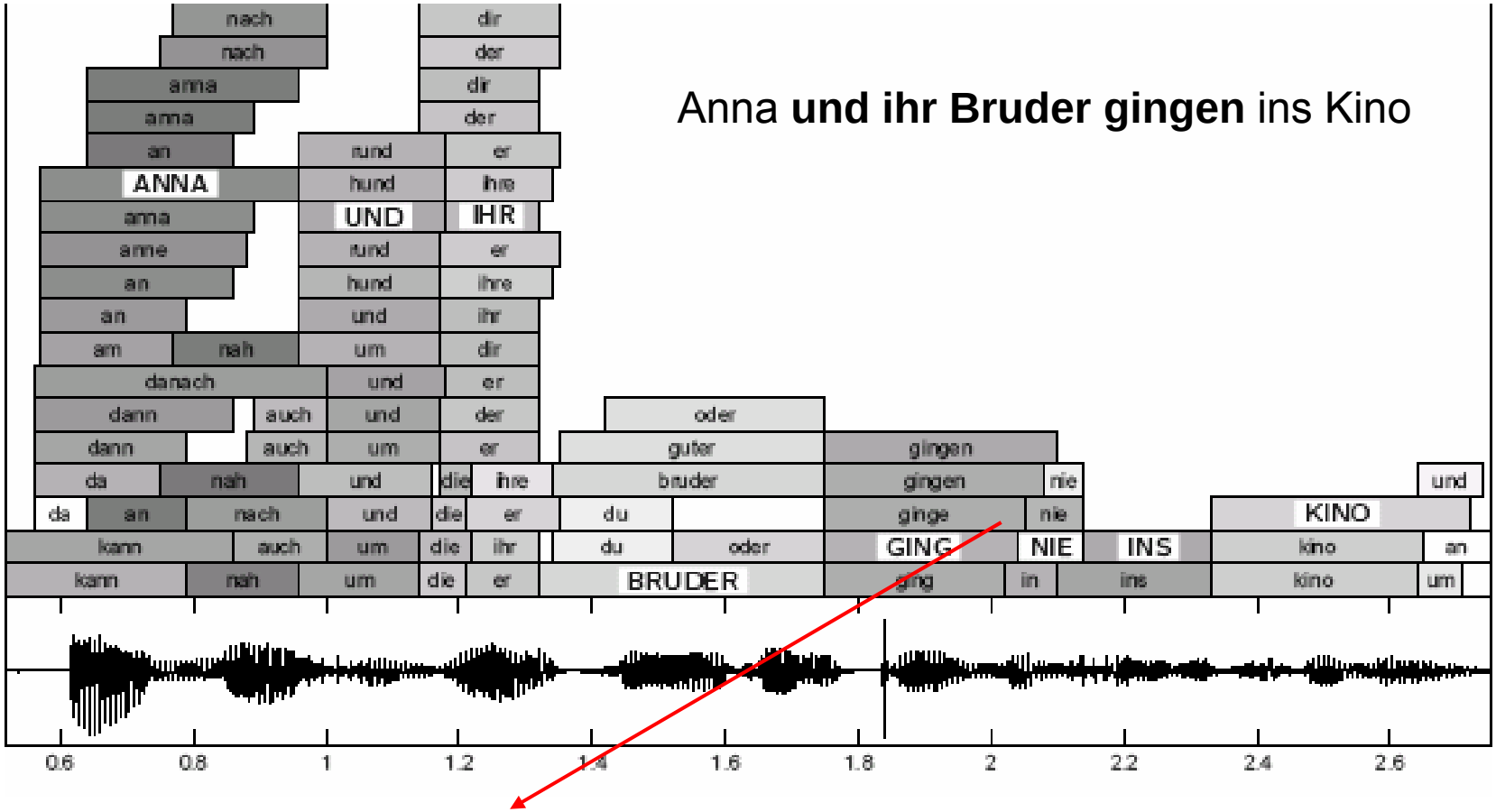
Stärken

- Die Modelle werden mit tatsächlich gebrauchten Sätzen trainiert. Die Modelle beschreiben also den wirklichen Gebrauch der Sprache und enthalten somit nicht nur syntaktische, sondern auch semantisches und pragmatisches Wissen.
- N-Grams können für eine beliebige Sprache mit den entsprechenden Daten (d.h. Texten dieser Sprache) trainiert werden. Sind diese Daten verfügbar, dann kann ein Sprachmodell erzeugt werden, und es ist dafür insbesondere kein linguistisches Wissen über die betreffende Sprache nötig.
- Nicht unwichtig ist schließlich, dass N-Grams einfach in den Erkenner integriert werden können. Man konfiguriert ein entsprechendes Erkennungsnetzwerk und wendet zur Erkennung den Viterbi-Algorithmus an. Die akustischen Modelle und das Sprachmodell werden also gleichzeitig eingesetzt, was sehr effizient ist.

Schwächen

- Modellschwäche
  - die statistische Sprachmodellierung kann aus praktischen Gründen nur innerhalb eines kurzen Kontextes die Zusammenhänge zwischen den Wörtern erfassen.
  - Das Beispiel illustriert diese Modellschwäche: Das Wortpaar “Bruder ging” kommt im Trainingstext viel häufiger vor als “Bruder gingen”. Das Sprachmodell greift deshalb zu kurz.

Beispiel – Bigrammodell



Anna und ihr Bruder ging nie ins Kino (syntaktisch nicht korrekt)

Anna schlief **und ihr Bruder ging** ins Kino → 4-Gram nicht ausreichend

Schwächen

- Überbewertung des Sprachmodells
  - Dies ist insbesondere dann der Fall, wenn die Viterbi-Wahrscheinlichkeit des akustischen Modells für das korrekte Wort zwar am größten, aber nur wenig größer als diejenige anderer Wörter ist und das Sprachmodell gleichzeitig dem korrekten Wort eine geringe A-priori-Wahrscheinlichkeit zuweist.

Beispiel

- Müssen z.B. in einer automatischen Telefonnummernauskunft nebst vielen andern die Namen “Möller” und “Müller” erkannt werden, dann wird im Unigram-Sprachmodell *P*(“Müller”) viel größer als *P*(“Möller”) sein.
- Das Sprachmodell wird somit bei der Erkennung der akustisch ähnlichen Namen “Müller” und “Möller” den ersteren stark begünstigen, was über die gesamte Einsatzzeit des Systems gesehen durchaus sinnvoll sein kann.
- Will jedoch eine Person tatsächlich die Telefonnummer eines Herrn Möller wissen, dann wird der Spracherkenner aufgrund des Sprachmodell mit sehr großer Wahrscheinlichkeit “Müller” erkennen.

Schwächen

- Trainingsdaten
- Um gute Resultate zu erhalten, müssen die Texte hinsichtlich Vokabular und Textart einigermaßen gut an die Anwendung angepasst sein. Ein passendes statistisches Sprachmodell zu erarbeiten, ist deshalb in der Praxis sehr aufwändig.