Grammatikmodelle

Vorlesungsunterlagen

Speech Communication 2, SS 2004

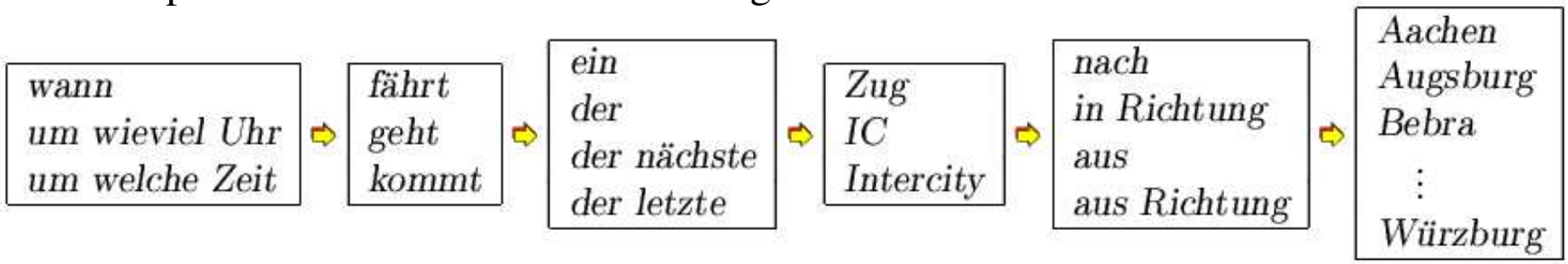
Franz Pernkopf/Erhard Rank

Institute of Signal Processing and Speech Communication
University of Technology Graz
Inffeldgasse 16c , 8010 Graz, Austria
Tel: +43 316 873 4431
E-Mail: pernkopf/rank@inw.tugraz.at

Typische Fehlentscheidungen bei der automatischen Spracherkennung:

(1)	„*den nächsten Zug*“	⇝	„*den nächste Zug*“	Kongruenzfehler
(2)	„*um sechs in Bonn*“	⇝	„*um sechs den Bonn*“	Vertauschung, Auslassung oder Einfügung von Funktionswörtern
(3)	„*von Essen nach Kiel*“	⇝	„*von ist in nach Kiel*“	Vertauschung von Inhaltswörtern durch ein oder mehr Funktionswörter
(4)	„*nach Köln fahren*“	⇝	„*nach Ulm fahren*“	grammatisch unauffällige Vertauschungen

Satzbauplan für einfache Benutzeräußerungen am Bahnhof



Bayes-Regel

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathcal{W}^*}{\operatorname{argmax}} P(\boldsymbol{w} \mid \boldsymbol{X})$$

mit den ***a posteriori Wahrscheinlichkeiten***

$$P(\boldsymbol{w} \mid \boldsymbol{X}) = \frac{\overbrace{P(\boldsymbol{X} \mid \boldsymbol{w})}^{\text{ASM}} \cdot \overbrace{P(\boldsymbol{w})}^{\text{LSM}}}{P(\boldsymbol{X})}$$

GESUCHT ist ein Wahrscheinlichkeitsmodell

$$P(w_1 w_2 \ldots w_T)$$

für Wortfolgen $\boldsymbol{w} = w_1 \ldots w_T$ der Länge T

Formel für bedingte Wahrscheinlichkeiten

$$P(w_1 w_2 \ldots w_T) = P(w_T \mid w_1 w_2 \ldots w_{T-1}) \cdot P(w_1 w_2 \ldots w_{T-1})$$

Kettenregel

$$
\begin{aligned}
P(w_1 w_2 \ldots w_T) &= P(w_1) \\
&\quad \cdot P(w_2 \mid w_1) \\
&\quad \cdot P(w_3 \mid w_1 w_2) \\
&\quad \cdot P(w_4 \mid w_1 w_2 w_3) \\
&\quad \cdot \ldots\ldots\ldots\ldots\ldots \\
&\quad \cdot P(w_T \mid w_1 \ldots w_{T-1}) \\
&= \prod_{t=1}^{T} P(w_t \mid w_1 \ldots w_{t-1})
\end{aligned}
$$

- $P(w_t \mid w_1 \ldots w_t)$ ist die ***bedingte*** Wortwahrscheinlichkeit
- die Wortkette $w_1 \ldots w_{t-1}$ heißt ***Vergangenheit*** oder ***Kontext*** des aktuellen Wortes w_t

Modelle mit beschränktem Kontext

$$P(w_t \mid w_1 \ldots w_{t-1}) \approx P(w_t \mid \underbrace{w_{t-n+1} \ldots w_{t-1}}_{(n-1)\text{-Gramm-Kontext}})$$

Unigramm-, Bigramm- und Trigramm-Modelle

$$P_{1G}(\boldsymbol{w}) = \prod_{t=1}^{T} P(w_t)$$

$$P_{2G}(\boldsymbol{w}) = P(w_1) \cdot \prod_{t=2}^{T} P(w_t \mid w_{t-1})$$

$$P_{3G}(\boldsymbol{w}) = P(w_1) \cdot P(w_2 \mid w_1) \cdot \prod_{t=3}^{T} P(w_t \mid w_{t-2} w_{t-1})$$

n-Gramm-Sprachmodelle

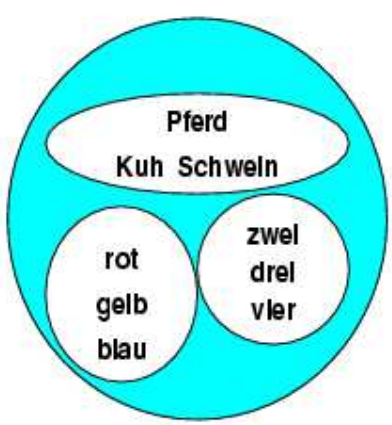
$$P_{nG}(\boldsymbol{w}) = \prod_{t=1}^{T} P(w_t \mid w_{t-n+1} \ldots w_{t-1})$$

mit $w_0, w_{-1}, w_{-2}, \ldots :=$ \$ (Satzanfangsmarkierungen)

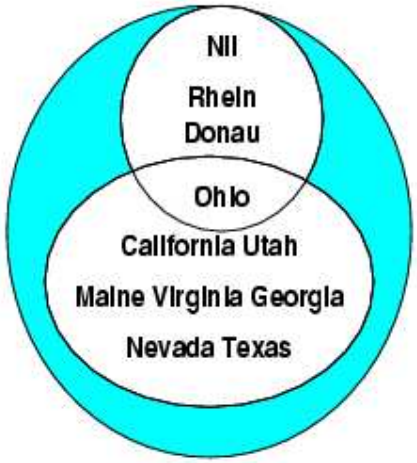
Gruppierung von Wörtern mit ähnlichem grammatischen u/o statistischen Verhalten

- ***syntaktisch:***
  Nomina, Verben, Adjektive von gleichem Genus, Kasus, Tempus, ...
- ***semantisch:***
  Ordinal- oder Kardinalzahlwörter;
  Wörter, die Farbe, Größe, Temperatur, ... charakterisieren;
- ***pragmatisch:***
  Eigennamen für Personen, Städte, Straßen, ... ;
  Nomina für Tiere, Pflanzen, Materialien, klinische Diagnosen, ...

Repräsentation von Wortassoziationen



disjunkte Kategorien



überlappende Kategorien

Maximum-Likelihood-Schätzung:

$$\hat{P}(v^i|\boldsymbol{v}) = \frac{\#(\boldsymbol{v}v^i)}{\#(\boldsymbol{v})} = \frac{\text{„Anzahl der } n\text{-Gramme } \boldsymbol{v}v^i\text{“}}{\text{„Anzahl der } (n-1)\text{-Gramme } \boldsymbol{v}\text{“}}$$

Problem beschränkten Stichprobenumfangs

- Wenn $\#(\boldsymbol{v}v^i) = 0$, so ist auch $\hat{P}(v^i|\boldsymbol{v}) = 0$
- Wenn $\#(\boldsymbol{v}) = 0$, so ist $\hat{P}(v^i|\boldsymbol{v})$ undefiniert!
- Insbesondere ist $\hat{P}(w) = 0$ für „neue“ oder “ungesehene“ Wörter

Glättung der Schätzwerte $\hat{P}$

Verfärbung der Zählfunktion

$$\hat{p}_k^{\mathrm{ML}} = \frac{N_k}{T} \Rightarrow \hat{p}_k = \frac{N_k^*}{T^*} = \frac{N_k^*}{\sum_{j=1}^{L} N_j^*}$$

Sobald $N_k^* > 0$ für alle k, so gilt auch $\hat{p}_k > 0$.

- **Laplace- oder uniforme Bayes-Schätzung**

$$N_k^* := N_k + 1 \Rightarrow \hat{p}_k = \frac{N_k + 1}{T + L}$$

- **Jeffrey-Glättung**

$$N_k^* := N_k + 1/2 \Rightarrow \hat{p}_k = \frac{N_k + 1/2}{T + L/2}$$

- **Quadratmittelmethode** (minimiert Fehlschätzungsrisiko)

$$N_k^* := N_k + 1/2\sqrt{T}$$

Beispiel 7.1 ($\boldsymbol{p} = (0.5, 0.4, 0.1)$ und $\boldsymbol{N} = (4, 3, 0)$)

- ML-Schätzung — $\hat{\boldsymbol{p}} = (0.57, 0.43, 0)$
- Laplace-Schätzung — $\hat{\boldsymbol{p}} = (0.5, 0.4, 0.1)$

Rückfallstrategien

Informationen in Sprachmodellen *niedrigerer Ordnung*:

$$\hat{P}(w_3 \mid w_1 w_2) \quad \Rightarrow \quad \hat{P}(w_3 \mid w_2) \quad \Rightarrow \quad \hat{P}(w_3) \quad \Rightarrow \quad {}^1\!/_L$$

Allgemeine Rückfallstrategie (Rekursionsgleichung)

$$\tilde{P}(w|\boldsymbol{v}) \;=\; \begin{cases} \hat{P}(w|\boldsymbol{v}) & \text{falls } \#(\boldsymbol{v}w) > 0 \\ \beta(\boldsymbol{v}) \cdot \tilde{P}(w|\boldsymbol{v}') & \text{falls } \#(\boldsymbol{v}w) = 0 \end{cases}$$

- $\hat{P}(w|\boldsymbol{v})$ ist ein Schätzer für $P(w|\boldsymbol{v})$ mit nichtverschwindenden Wahrscheinlichkeiten.
- $\boldsymbol{v}'$ ist der abgemagerte Bedingungsteil des gröberen Modells.

Interpolierte n-Gramme

$$\begin{aligned}\tilde{P}^{(n)}(w_t \mid w_1 \ldots w_{t-1}) = \lambda_0^{(n)} \cdot \frac{1}{L} &+ \lambda_1^{(n)} \cdot \hat{P}(w_t) \\ &+ \lambda_2^{(n)} \cdot \hat{P}(w_t|w_{t-1}) \\ &+ \sum_{i=3}^{n} \lambda_i^{(n)} \cdot \hat{P}(w_t \mid w_{t-i+1} \ldots w_{t-1})\end{aligned}$$

Interpolierte n-Gramme

$$\begin{aligned}\tilde{P}^{(n)}(w_t \mid w_1 \ldots w_{t-1}) \;=\; & \lambda_0^{(n)} \cdot \frac{1}{L} \;+\; \lambda_1^{(n)} \cdot \hat{P}(w_t) \\ & +\; \lambda_2^{(n)} \cdot \hat{P}(w_t | w_{t-1}) \\ & +\; \sum_{i=3}^{n} \lambda_i^{(n)} \cdot \hat{P}(w_t \mid w_{t-i+1} \ldots w_{t-1})\end{aligned}$$

- Falls $t < n$, fülle $w_{t-n+1}, \ldots, w_0$ mit \$-Marken auf:

$$\tilde{P}^{(n)}(\boldsymbol{w}) \;=\; \prod_{t=1}^{T} \tilde{P}^{(n)}(w_t \mid w_1 \ldots w_{t-1})$$

- Falls $t < n$, verwende $\tilde{P}^{(t)}$ oder $\tilde{P}^{(t+1)}$ statt $\tilde{P}^{(n)}$:

$$\tilde{P}^{(n)}(\boldsymbol{w}) \;=\; \prod_{t=1}^{n} \tilde{P}^{(t)}(w_t \mid w_1 \ldots w_{t-1}) \;\cdot\; \prod_{t=n+1}^{T} \tilde{P}^{(n)}(w_t \mid w_1 \ldots w_{t-1})$$

- Im Falle $\#(\boldsymbol{w}_{t-i+1}^{t-1}) = 0$ wird $\hat{P}(w_t \mid \boldsymbol{w}_{t-i+1}^{t-1}) = 1/L$ substituiert.