

# Softwarepraktikum „Muster- und Bilderkennung“ im Grundstudium

## Aufgabe 4 Sprachmodellierung und mehrdeutige Tastaturen

Saša Hasan, Evgeny Matusov  
Human Language Technology and Pattern Recognition  
Lehrstuhl für Informatik 6  
RWTH Aachen University

# 1 Sprachmodellierung

- Sprachmodelle bewerten die syntaktische Korrektheit eines Satzes

$$w_1 w_2 \dots w_N = w_1^N$$

$$P(w_1^N) = P(w_1)P(w_2|w_1) \dots P(w_N|w_1^{N-1})$$

$$= \prod_{i=1}^N P(w_i|w_1^{i-1})$$

$$\approx \prod_{i=1}^N P(w_i|w_{i-n+1}^{i-1}) \quad \text{Beschränkung auf lokalen Kontext}$$

- lokaler Kontext wird als  $n$ -Gramm bezeichnet:

|          | $n$ | Beobachtung         | Wahrscheinlichkeit      |
|----------|-----|---------------------|-------------------------|
| Unigramm | 1   | $w_i$               | $P(w_i)$                |
| Bigramm  | 2   | $w_{i-1}w_i$        | $P(w_i w_{i-1})$        |
| Trigramm | 3   | $w_{i-2}w_{i-1}w_i$ | $P(w_i w_{i-2}w_{i-1})$ |

## Beispiel: Bigramme für den Satz

|   |  |                              |
|---|--|------------------------------|
| <code>&lt;s&gt; dies ist ein Haus &lt;/s&gt;</code> |  |                              |
| <hr/>   |  |                              |
| <code>&lt;s&gt; dies</code>                         |  | $P(\text{dies} \text{<s>})$  |
| <code>  dies ist</code>                             |  | $P(\text{ist} \text{dies})$  |
| <code>    ist ein</code>                            |  | $P(\text{ein} \text{ist})$   |
| <code>      ein Haus</code>                         |  | $P(\text{Haus} \text{ein})$  |
| <code>        Haus &lt;/s&gt;</code>                |  | $P(\text{</s>} \text{Haus})$ |

Wahrscheinlichkeiten werden auf Trainingsdaten geschätzt

Problem: Was passiert mit *ungesehenen*  $n$ -Grammen?

Beispiel: `dies war ein Haus`

$P(\text{war}|\text{dies})$  sei ein vorher nicht gesehenes Bigramm. Was nun?

⇒ Backoff-Modelle

## Backoff-Modell für Bigramme:

$$P(w_i|w_{i-1}) = \begin{cases} P(w_i|w_{i-1}) & \text{für vorhandenen Eintrag } w_{i-1} \ w_i \\ \alpha(w_{i-1}) \cdot P(w_i) & \text{sonst} \end{cases}$$

$\alpha(w_{i-1})$  ist hierbei ein Normalisierungsfaktor, so dass  $P(w_i|w_{i-1})$  normiert ist, d.h.  $\sum_w P(w|h) = 1$

vorheriges Beispiel: unbekanntes Bigramm dies war

Backoff zum Unigramm mit Gewichtung des Kontextes:  $\alpha(\text{dies})P(\text{war})$

# ARPA-Format für Sprachmodelle:

\data\  
 ngram 1=6998  
 ngram 2=33978  
 ngram 3=6283

\1-grams:

-1.777618          a                  -0.3105851

-4.038711          a.m.              -0.4448084

-4.184839          abandon          -0.1696519

...

$\log_{10} P(w_i)$

$w_i$

$\alpha(w_i)$

\2-grams:

-2.799451          a chair          0.1895401

-3.089627          a chance

-2.670388          a charge        0.06958468

...

$\log_{10} P(w_i|w_{i-1})$

$w_{i-1} w_i$

$\alpha(w_{i-1}w_i)$

\3-grams:

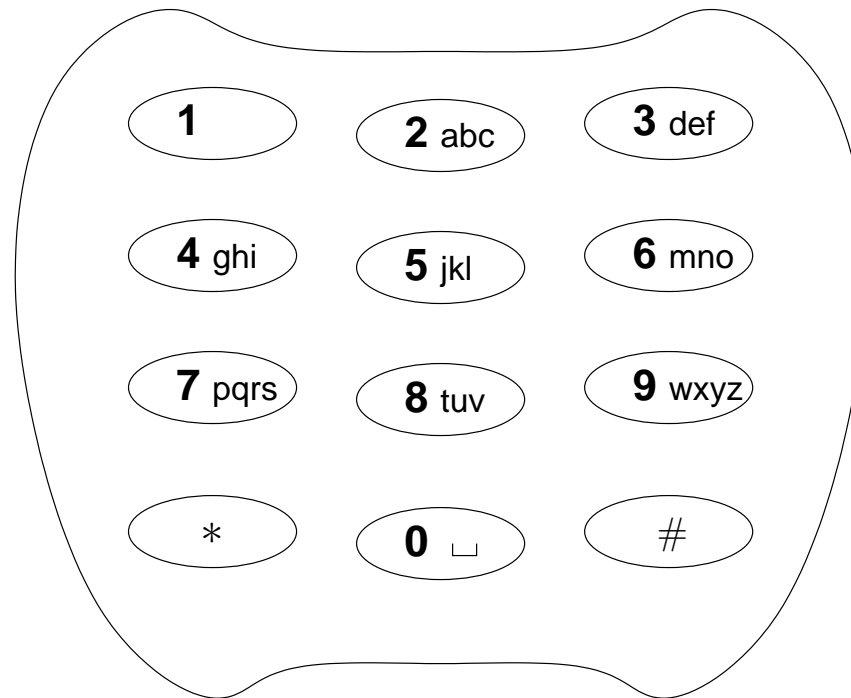
-0.4819637          where are we

-1.161494          you are a

-1.385054          you are going

...

## 2 Mehrdeutige Tastaturen



**Exakte Matches:**

**7 3 2 5**  $\implies$  **real  
peak  
seal**

**mit Vervollständigung:**

**7 3 2 5**  $\implies$  **really  
realise  
reality  
reckon  
reckoning  
secluded**

## Tippen von mehrdeutigen Texten:

4 8 4 7 8 4 6 3 8 6 ...  
 it is time to ...

Mit Hilfe des Sprachmodells werden nun alle passenden Vorschläge zu einem Code bewertet und die Vorschlagsliste dementsprechend sortiert.

| Disambiguiert   | Vorschläge $w$  | $P(w is)$  |
|---|---|--|
| <span style="border: 1px solid black; padding: 2px;">4</span> <span style="border: 1px solid black; padding: 2px;">8</span> <span style="border: 1px solid black; padding: 2px;">4</span> <span style="border: 1px solid black; padding: 2px;">7</span><br>it            is | <span style="border: 1px solid black; padding: 2px;">8</span><br>true<br>the<br>to<br>that<br>this<br>... | <b>0.05882</b><br><b>0.04071</b><br><b>0.00740</b><br><b>0.00740</b><br><b>0.00550</b> |

| Disambiguiert                         | Vorschläge $w$     | $P(w is)$      |
|---------------------------------------|--------------------|----------------|
| <b>4 8</b> <b>4 7</b><br>it        is | <b>8 4</b><br>that | <b>0.00740</b> |
|                                       | this               | <b>0.00550</b> |
|                                       | they               | <b>0.00149</b> |
|                                       | there              | <b>0.00092</b> |
|                                       | time               | <b>0.00077</b> |
|                                       | ...                |                |

| Disambiguiert                         | Vorschläge $w$       | $P(w is)$      |
|---------------------------------------|----------------------|----------------|
| <b>4 8</b> <b>4 7</b><br>it        is | <b>8 4 6</b><br>time | <b>0.00078</b> |
|                                       | thousand             | <b>0.00053</b> |
|                                       | thought              | <b>0.00038</b> |
|                                       | though               | <b>0.00021</b> |
|                                       | those                | <b>0.00019</b> |
|                                       | ...                  |                |