

Softwarepraktikum „Muster- und Bilderkennung“ im Grundstudium

Aufgabe 6 Sprachmodellierung und mehrdeutige Tastaturen

**Saša Hasan, Evgeny Matusov
Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
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><s> dies ist ein Haus </s></code>		
<hr/>		
<code><s> 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 </s></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:

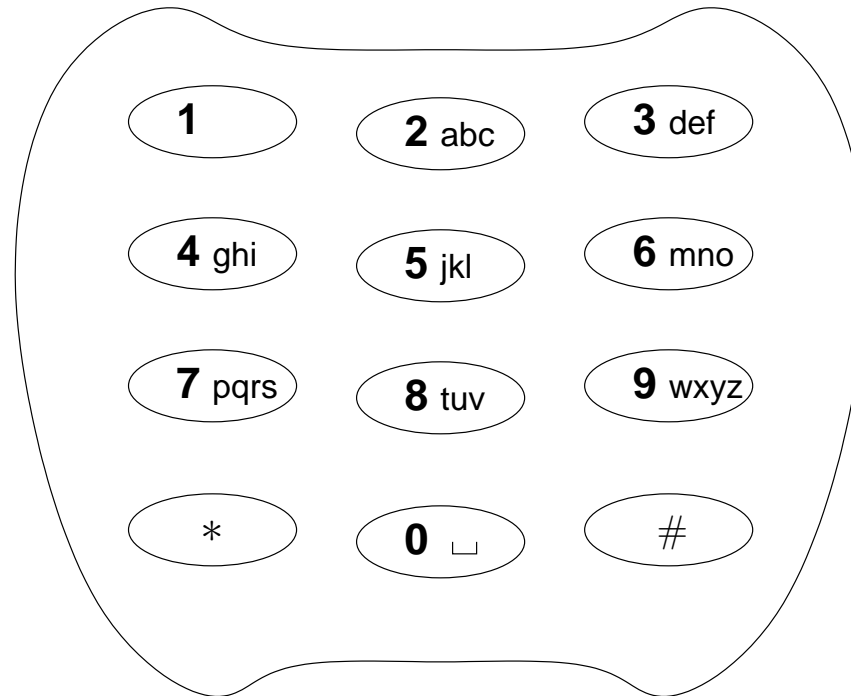
-2.799451	a chair	0.1895401
-3.089627	a chance	
-2.670388	a charge	0.06958468
...		

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

\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	$P(w is)$															
<table border="1"> <tr> <td>4</td><td>8</td><td>4</td><td>7</td> </tr> <tr> <td>it</td><td>is</td><td></td><td></td> </tr> </table>	4	8	4	7	it	is			<table border="1"> <tr> <td>8</td> </tr> <tr> <td>true</td> </tr> <tr> <td>the</td> </tr> <tr> <td>to</td> </tr> <tr> <td>that</td> </tr> <tr> <td>this</td> </tr> <tr> <td>...</td> </tr> </table>	8	true	the	to	that	this	...	<p>0.05882</p> <p>0.04071</p> <p>0.00740</p> <p>0.00740</p> <p>0.00550</p>
4	8	4	7														
it	is																
8																	
true																	
the																	
to																	
that																	
this																	
...																	

Disambiguiert	Vorschläge w	$P(w is)$
4 8 4 7 it is	8 4 that	0.00740
	this	0.00550
	they	0.00149
	there	0.00092
	time	0.00077
	...	

Disambiguiert	Vorschläge w	$P(w is)$
4 8 4 7 it is	8 4 6 time	0.00078
	thousand	0.00053
	thought	0.00038
	though	0.00021
	those	0.00019
	...	