Gegeben sind die Buchstaben-N-Gramm-Häufigkeiten f(a,a)=1, f(a,b)=2, f(b,a)=0 und f(b,b)=1.

Berechnen Sie den Discount für die Glättung der Bigramm-Wahrscheinlichkeiten. (1 Punkt)

Berechnen Sie dann die Unigramm-Häufigkeiten f(a) und f(b) einmal nach dem Standard-Backoff-Verfahren und einmal nach dem Kneser-Ney-Verfahren. (1 Punkt)

Berechnen Sie aus den Kneser-Ney-Häufigkeiten die ungeglätteten Unigramm-Wahrscheinlichkeiten p(a) und p(b). (1 Punkt)

Discount

N1 = 2
N2 = 1
Discount_bigram = N1 /N1+ (2*N2) = 2 / 2 + (2 * 1) = 2 / 4 = 0.5

Unigram-Häufigkeiten

Standard-Backoff-Verfahren

f(a) = f(a,a) +f(b,a) = 1 +0 = 1
f(b) = f(a,b) +f(b,b) = 2+ 1 = 3
p(a) = f(a) / f(a) + f(b) = 1 / 1 +3
p(b) = f(b) / f(a) + f(b) = 3 / 1+3

Kneser-Ney-Verfahren

f*(a) = 1(a,a) +1(b,a) = 1 + 0 = 1
f*(b) = 1(a,b) +1(b,b) = 1 + 1 = 2
p(a) = f*(a) / f*(a) + f*(b) = 1 / 1 +2 = 1/3
p(b) = f*(b) / f*(a) + f*(b) = 2 / 1+2 = 2 /3

Kneser-Ney Backoff-Verteilung

Bei der bisherigen Berechnung der n-1-Gramm-Häufigkeiten zur Schätzung der Backoff-Verteilung summieren wir die Häufigkeiten über alle möglichen Vorgängerwörter w':

$$f(C, w) = \sum_{w'} f(w', C, w)$$

C ist eine (eventuell leere) Folge von Wörtern.

Bei Kneser-Ney zählen wir, wieviele **unterschiedliche** Wörter vor dem Wort-n-Gramm aufgetreten sind:

$$f^*(C, w) = \sum_{w'} \mathbf{1}_{f(w', C, w) > 0}$$

$\mathbf{1}_{test}$ ist 1, falls *test* wahr ist und sonst 0.

Die Kneser-Ney-Methode zählt n-Gramm-Types (statt -Tokens).

Aus den so ermittelten Häufigkeiten, werden dann die Parameter der Backoff-Wahrscheinlichkeits-Verteilungen geschätzt.

$$p_{backoff}(w|C) = \frac{f^*(C, w)}{\sum_{w'} f^*(C, w')}$$