3. Übung zur Vorlesung "NLP – Analyse des Wissensrohstoffes Text" im Sommersemester 2008

Dr. Andreas Hotho, Dipl.-Inform. Dominik Benz, Wi.-Inf. Beate Krause

14. Mai 2008

1 Kollokationen

Gegeben ist ein finnischer Korpus mit $N = 28181344$ unterschiedlichen Wörtern. Tabelle 1 gibt die Häufigkeiten einzelner Wörter, so wie die Häufigkeiten des gemeinsamen Auftretens an.

1. Berechnen Sie den Mittelwert und die Varianz für die Wortpaare aus der Tabelle. Beurteilen Sie, wie gut diese Methode für die Entdeckung von Kollokationen geeignet ist. Wie beeinflusst die Fenstergröße das Ergebnis?
2. Benutzen Sie den t-test und Pearson's Chi-square Test für das Wortpaar "valkoinen" und "talo". Berücksichtigen Sie dabei nur $C(w_1, w_2)$ für das gemeinsame Aufkommen. Formulieren Sie eine Nullhypothese. Nutzen Sie das Signifikanzniveau von 0.05. Welche unterschiedlichen Annahmen treffen die beiden Tests?

Übersetzung der Worte:

- hakea = apply for, työ = job

Tabelle 1: Häufigkeiten für das Auftreten einzelner Worte und der Kombination von Wortpaaren

w_1	w_2	C(w_1)	C(w_2)	C(w_1,w_2)	C(w_1,x,w_2)	C(w_2,w_1)	C(w_2,x,w_1)
hakea	työ	10435	26174	31	26	22	11
valkoinen	talo	3665	10767	710	2	1	6
herne	nenä	115	974	3	0	0	0
ja	olla	818046	1387476	7329	39979	3612	38162
venäjä	presidentti	27637	26855	717	216	10	24

- valkoinen = white, talo = house
- herne = pea, nenä = nose, “herne nenä” = “pissed of”
- ja = and, olla = be
- Venäjä = Russia, presidentti = president

2 n-Gramme

Gegeben ist folgender Text.

```
JOHN READ MOBY DICK
MARY READ A DIFFERENT BOOK
SHE READ A BOOK BY CHE
```

1. Berechnen Sie die Wahrscheinlichkeit für die Sequenzen “John read a book” und “Cher read a book” mit der Maximum Likelihood Estimate Methode. Betrachten Sie dabei die aufeinanderfolgenden Wortpaare (Bigramme) und multiplizieren Sie die Gesamtwahrscheinlichkeit. Erläutern Sie anhand Ihrer Ergebnisse, warum Smoothing für die Berechnung von n-Grammen wichtig ist.
2. Berechnen Sie für die gleichen Sätze die Maximum Likelihood Wahrscheinlichkeiten mit Laplace und Lidstone Smoothing. Setzen Sie beim Lidstone Smoothing $\lambda = 0.5$.
3. Welche Markov-Eigenschaft spiegelt sich bei den berechneten Bigrammen wider?
4. Zusatzaufgabe: Welche weiteren Smoothing-Methoden könnte man auf dieses Beispiel anwenden?

3 Markov Modelle

In den letzten Monaten haben Sie Wetterwechsel beobachtet indem Sie jeden Tag um die gleiche Uhrzeit notiert haben, ob es sonnig (S_1), wolkig (S_2) oder regnerisch (S_3) ist. Dabei sind Sie auf die folgenden Beobachtungen für Wetterwechsel gekommen:

$$A = \begin{pmatrix} 0.8 & 0.15 & 0.05 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

Die Matrix A stellt die Übergangswahrscheinlichkeiten zwischen den einzelnen Zuständen dar. Zum Beispiel wissen wir, dass die Wahrscheinlichkeit, dass es, wenn es heute sonnig ist, morgen wolkig sein wird, 15% beträgt.

1. Zeichnen Sie eine Markov Kette mit den zugehörigen Wahrscheinlichkeiten. Handelt es sich um eine versteckte oder sichtbare Markovkette?
2. Heute ist es wolkig. Was ist die Wahrscheinlichkeit für das folgende 5-Tage-Wetter: regnerisch - wolkig - sonnig - sonnig - sonnig.

Stellen Sie sich jetzt vor, dass Ihr kolumbianischer Freund folgende Sportarten betreibt: schwimmen, laufen und radeln. Die tägliche Sportart entscheidet sich nach dem Wetter. In Kolumbien scheint entweder die Sonne oder es regnet.

- Wenn heute die Sonne scheint, scheint mit 70% morgen wieder die Sonne
- Wenn es regnet, regnet es am nächsten Tag wieder mit einer Wahrscheinlichkeit von 35%.

Dabei geht Ihr Freund an einem sonnigen Tag mit 60% Wahrscheinlichkeit radeln und mit 30% laufen. An einem regnerischen Tag geht er mit 80% Wahrscheinlichkeit schwimmen, ansonsten entweder laufen oder radeln. Am Telefon erzählt ihr Freund regelmäßig, welche Sportart er am aktuellen Tag betrieben hat.

1. Geben Sie die Übergangsmatrix A und die Ausgabewahrscheinlichkeiten des Hidden Markov Models an.
2. Bestimmen Sie die Wahrscheinlichkeit, dass Ihr Freund in den nächsten zwei Tagen erst radeln und dann schwimmen wird. Heute scheint die Sonne. Mit welchem Algorithmus kann man dieses Problem effizient lösen?
3. Mit welchem Algorithmus lässt sich die folgende Problemstellung lösen? Welches Wetter war in Kolumbien am Wahrscheinlichsten, wenn ihr Freund in den letzten zwei Tagen erst radeln, dann schwimmen war?

4 Praxisübung - Hidden Markov Modelle (Abgabe 27. Mai)

Nehmen Sie das zweite Beispiel der letzten Aufgabe (den kolumbianischen Sportler) und implementieren Sie

- den Forward Algorithmus um die Wahrscheinlichkeiten von Sportaktivitäten für die nächste Woche zu berechnen. Gehen Sie davon aus, dass zu Beginn immer die Sonne scheint.
- den Viterbi Algorithmus um für eine gegebene wöchentliche Ausgabesequenz das wahrscheinlichste Wetter zu ermitteln.