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

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?

1. Wir definieren die folgenden Wortabstände: -1, wenn beide Wörter direkt folgen, -2 wenn ein Wort dazwischen auftritt, 1 wenn die beiden Wörter umgekehrt stehen,

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

Tabelle 2: Mittelwerte und Varianz für die Wortpaare

w_1	w_2	Mean	Varianz
hakea	työ	-0.433	2.068
valkoinen	talo	-0.975	0.091
herne	nenä	-1.00	0.00
ja	olla	-0.083	3.625
venäjä	presidentti	-1.128	0.472

und 2 wenn die beiden Wörter umgekehrt stehen und ein Wort dazwischen steht. Für die Kollokation "valkoinen" und "talo" gilt:

$$\begin{aligned}
\text{Mean(valkoinen, talo)} &= \frac{-1 * 710 - 2 * 2 + 1 + 2 * 6}{710 + 2 + 1 + 6} \\
\text{Var(valkoinen, talo)} &= \frac{(-1 - (-0.975))^2 * 710 + (-2 - (-0.975))^2 * 2}{89} \\
&\quad + \frac{(1 - (-0.975))^2 * 1 + (2 - (-0.975))^2 * 6}{89} \\
&\approx 0.083
\end{aligned}$$

Weiter Lösungen: Die Wortpaare können beginnend mit der niedrigsten Varianz angeordnet werden. Bei Wortpaaren mit wenigen Vorkommen liefert die Methode allerdings noch keine guten Ergebnisse. Auch Wortpaare, die sehr häufig vorkommen, werden sehr gut bewertet.

Die Größe des Fensters spielt eine wichtige Rolle. Wenn es zu groß ist, können Wortpaare zufällig zu häufig auftreten, wenn es zu klein ist, findet man keine Kollokationen die weiter auseinander stehen.

2. Unsere Nullhypothese besagt, dass die Worte in einem Paar unabhängig sind: $\mathrm{P}(s_1, s_2) = \mathrm{P}(S_1)\mathrm{P}(S_2)$. Im t-Test nehmen wir an, dass die Wahrscheinlichkeiten normalverteilt sind und prüfen, ob die Erwartungswerte der beobachteten Daten von den Erwartungswerten der Nullhypothese abweichen. Die t-Werte sind gegeben durch

$$t = \frac{\hat{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

μ ist der Erwartungswert der Verteilung, also

$$\mu = P(S_1)P(S_2) = \frac{C(s_1)}{N}\frac{C(s_2)}{N}.$$

Tabelle 3: nxn Tabelle zur Berechnung der Erscheinungshäufigkeiten für den Chi-Square-Test

	w_1 = valkoinen	$w_1 \neq$ valkoinen
w_2=talo	710	10767 - 710
$w_2 \neq$ talo	3665 - 710	28181344 - 710 - 10057 - 2955

$\hat{x}$ ist der Erwartungswert für ein gemeinsames Auftreten der beiden Worte aus dem vorliegenden Text, also

$$\hat{x} = \frac{P(s_1, s_2)}{N} = p.$$

s^2 ist die Varianz des mit $p(1-p) \approx p$ (Vergleiche dafür die Bernoulli-Test Beschreibungen im Buch).

Für das Paar "valkoinen" und "talo" ergibt sich:

$$t = \frac{\frac{710}{28181344} - \frac{3665*10767}{28181344^2}}{\sqrt{\frac{\frac{710}{28181344}}{28181344}}} \approx 27$$

Wenn der t-Wert über 6.314 liegt (dies kann man in einer Tabelle nachschlagen), dann ist die Wahrscheinlichkeit, dass das Sample von der Verteilung aus der Nullhypothese (also aus einer unabhängigen Verteilung) kleiner als 5%. In diesem Fall kann die Nullhypothese also abgelehnt werden.

3. Im Gegensatz zum t-Test wird im χ^2-Test keine Normalverteilung angenommen. Der χ^2-Test schaut auf die getrennten Wahrscheinlichkeiten der beiden Wörter und schätzt wie häufig die beiden Wörter miteinander auftreten sollten. Dieser Wert wird mit dem beobachteten Wert verglichen; falls die Wert sich genügend stark unterscheiden, ist es wahrscheinlich, dass das Wortpaar eine Kollokation ist. Die $2x2$ Tabelle **??** für das Wortpaar ist Wenn man die Formel (5.7) auf S. 170 anwendet, erhält man die folgende Gleichung:

$$\chi^2 = \frac{28181344(710 * 28167622 - 10057 * 2955)}{(710 + 10057)(710 + 2955)(10056 + 28167622)(2955 + 28167622)} \approx 358771$$

Wenn der χ^2-Text einen Wert über den kritischen Wert von 3.841 (bei einem Signifikanzniveau von 5%) ergibt, dann können wir die Nullhypothese, dass die beiden Wörter unabhängig voneinander auftauchen, ablehnen. Dies ist hier der Fall.

Übersetzung der Worte:

- hakea = apply for, työ = job
- 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.

   $$\begin{aligned}
   &\text{P(John Read A Book)}\\
   &= \text{p(John|*)p(Read|John)p(A|Read)p(Book|A)p(*|Book)}\\
   &= \frac{\text{c(* John)}}{\sum_w \text{c(* w)}} \frac{\text{c(John Read)}}{\sum_w \text{c(John w)}} \frac{\text{c(Read A)}}{\sum_w \text{c(Read w)}} \frac{\text{c(A Book)}}{\sum_w \text{c(A w)}} \frac{\text{c(Book *)}}{\sum_w \text{c(Book w)}}\\
   &= \frac{1}{3}\frac{1}{1}\frac{2}{3}\frac{1}{2}\frac{1}{2}\\
   &\approx 0.06
   \end{aligned}$$

   $$\begin{aligned}
   &\text{P(Cher Read A Book)}\\
   &= \text{p(Cher|*)p(Read|Cher)p(A|Read)p(Book|A)p(*|Book)}\\
   &= \frac{\text{c(* Cher)}}{\sum_w \text{c(* w)}} \frac{\text{c(Cher Read)}}{\sum_w \text{c(Cher w)}} \frac{\text{c(Read A)}}{\sum_w \text{c(Read w)}} \frac{\text{c(A Book)}}{\sum_w \text{c(A w)}} \frac{\text{c(Book *)}}{\sum_w \text{c(Book w)}}\\
   &= \frac{0}{3}\frac{0}{1}\frac{2}{3}\frac{1}{2}\frac{1}{2}\\
   &= 0
   \end{aligned}$$

   Während einige Events sehr häufig auftreten, gibt es unendlich viele seltenene Events, die nicht, oder nur sehr selten in einem Datensatz enthalten sind (Zipf-Verteilung). Zum Beispiel kommt das Event p(Read|Cher) nicht vor. Genauso schwierig wäre ein Satz mit einem neuen Wort. Diese Problematik lässt sich mit der Maximum Likelihood Methode nicht darstellen, die Wahrscheinlichkeiten werden 0.

2. Berechnen Sie für die gleichen Sätze die Maximum Likelihood Wahrscheinlichkeiten mit Laplace und Lidstone Smoothing. Setzen Sie beim Lidstone Smoothing $\lambda = 0.5$.

   Laplace Smoothing:

P(John Read A Book)
$$= \frac{1+1}{11+3}\frac{1+1}{11+1}\frac{1+2}{11+3}\frac{1+1}{11+2}\frac{1+1}{11+2}$$
$$\approx 0.0001$$

P(Cher Read A Book)
$$= \frac{1+0}{11+3}\frac{1+0}{11+1}\frac{1+2}{11+3}\frac{1+1}{11+2}\frac{1+1}{11+2}$$
$$\approx 0.00003$$

Lidstone's Law:

P(John Read A Book)
$$= \frac{0.5+1}{11+0.5*3}\frac{0.5+1}{11+0.5*1}\frac{0.5+2}{11+0.5*3}\frac{0.5+1}{11+0.5*2}\frac{0.5+1}{11+0.5*2}$$
$$\approx 0.7$$

P(Cher Read A Book)
$$= \frac{0.5+0}{11+0.5*3}\frac{0.5+0}{11+0.5*1}\frac{0.5+2}{11+0.5*3}\frac{0.5+1}{11+0.5*2}\frac{0.5+1}{11+0.5*2}$$
$$\approx 0.533$$

3. Welche Markov-Eigenschaft spiegelt sich bei den berechneten Bigrammen wider? Die nachfolgenden Zustände (Wörter) sind nur vom aktuellen (hier von einem, sonst von n) Vorgänger abhängig.
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.

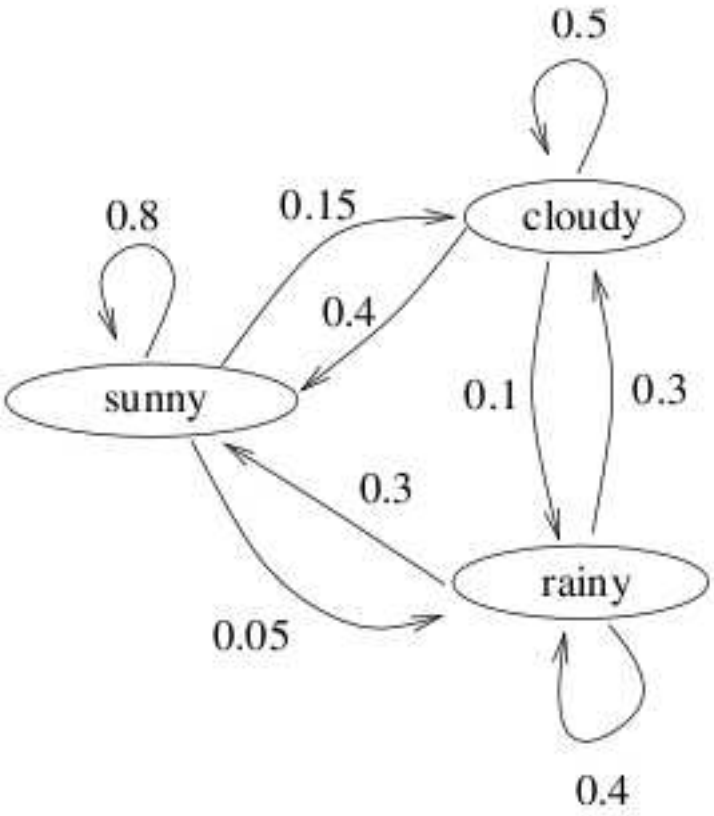

Abbildung 1: Zustandsautomat für die Transitionen der Matrix A aus 3.1

1. Zeichnen Sie eine Markov Kette mit den zugehörigen Wahrscheinlichkeiten. Handelt es sich um eine versteckte oder sichtbare Markovkette?

   Die Markovkette ist eine sichtbare Kette, da es keine versteckten Zustände gibt.

2. Heute ist es wolkig. Was ist die Wahrscheinlichkeit für das folgende 5-Tage-Wetter: regnerisch - wolkig - sonnig - sonnig - sonnig.

   Gesucht ist die Wahrscheinlichkeit für die Sequenz S, gegeben das der aktuelle Zustand S_2 ist.

$$\begin{aligned} P(S|q_0 = S_2) &= a_{23}a_{32}a_{21}a_{11}a_{11} \\ &= 0.1 * 0.3 * 0.4 * 0.8 * 0.8 \\ &= 0.0077 \end{aligned}$$

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.

	Schwimmen	Radeln	Laufen
Sonne	0.1	0.6	0.3
Regen	0.8	0.1	0.1

- 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.

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.65 & 0.35 \end{pmatrix}$$

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? Es müssen alle möglichen Pfade aufaddiert werden. Dies ergibt:

   $0.7*0.6*0.7*0.1+0.7*0.6*0.3*0.1+0.3*0.6*0.35*0.8+0.3*0.6*0.65*0.8 = 0.0294+0.0126+0.0504+0.093$

   Der geeignete Algorithmus ist der Forward Algorithmus.

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?

   Dies lässt sich mit dem Viterbi-Algorithmus lösen.

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.