Klausur 2009: Quantitative Methoden der CL

1 Nennen Sie Vor- und Nachteile statistischer Verfahren (im Gegensatz zu nichtstatistischen Ansätzen)

Vorteile: Ranking von Ergebnissen, meist robuster, basierend auf Empirie, schnell
Nachteile: Modelle meist nicht interpretierbar, annotierte Daten nötig

2 Schätzen von Wahrscheinlichkeiten

In einem Corpus mit 1000 Wörtern wird das Tag A 100 mal vergeben und das Tag B 50 mal. Das letzte Wort im Corpus ist weder B noch A. Das Bigram 'A B' kommt 30 mal vor.
a) Schätzen Sie folgende Wahrscheinlichkeiten anhand MLE:

- P(A) und P(B): P(A)= 100/1000, P(B) = 50/1000
- P(B|A): 30/100 (gemeinsam/Bigramme mit A)
- P(A,B): 30/999 (gemeinsam/alle Bigramme)

b) Berechnen Sie P(A|B) anhand der Bayes'schen Formel.
P(A|B) = (P(A) * P(B|A))/P(B) = (1/10 * 3/10)/(5/100)=3/5
c) Sind P(A) und P(B) stochastisch unabhängig? Belegen Sie Ihre Aussage mathematisch.
Nein: P(A,B) ist ungleich P(A) * P(B)

3 Präzision, Ausbeute

Ihr Tagger hat 160 Wortformen als Adjektive getaggt. Laut Gold Standard gibt es aber 200 Adjektive. 100 Tagger-Entscheidungen sind richtig. Berechnen Sie Präzision und Ausbeute.
Präzision: 100/160 Ausbeute: 100/200

4 Witten-Bell Smoothing

Die Wahrscheinlichkeitsmasse für nicht-gesehene Ereignisse ist:

$$\frac{T}{N+T}$$

mit T=Anzahl der beobachteten Types und N=Anzahl der Token Verteilt auf alle nicht-gesehenen Ereignisse ergibt das:

$$\frac{T}{Z(N+T)}$$

wobei Z die Menge der Nullereignisse ist.

Alle *beobachteten* Ereignisse haben die Schätzung (c_i als Zählfunktion):

$$\frac{c_i}{N+T}$$

Aufgabe: Wahrscheinlichkeit von Lesarten schätzen. Das Wort 'love' hat in WordNet 6 Nomenlesarten. Sie möchten die Wahrscheinlichkeit für diese Lesarten schätzen. Sie machen eine Stichprobe - Sie disambiguieren von Hand 100 Vorkommen des Wortes 'love' in einem Corpus. Nehmen wir der Einfachheit halber an, dass 4 der 6 Lesarten jeweils 25 mal auftreten, 2 Lesarten also nicht angetroffen werden. Berechnen Sie a) nach Witten-Bell die Wahrscheinlichkeit eines einzelnen nicht-gesehenen Ereignisses. Was ist b) die Wahrscheinlichkeit (jeweils) der vier gesehenen Ereignisse?
Masse für nicht-gesehene: 4/(100+4)
Masse pro nicht-gesehenes: 4/(2*(100+4))= 1/52
gesehene: 25/104

5 Parsing

Gegeben folgende Grammatik. Was ist die Wahrscheinlichkeit, das folgender Satz von der Grammatik generiert wird: 'Theo liebt Anna'? Sie brauchen den konkreten Zahlenwert nicht zu ermitteln, nur die Berechnungsvorschrift.

```
0.8  S -> NP VP
0.2  S -> V NP
```

```
0.3  NP -> PN
0.8  NP -> Det Noun

0.4  VP -> V
0.6  VP -> V NP

0.1 PN -> Anna
0.1 PN -> Theo
..

0.25 Verb -> liebt
...
```

Lösung: 2* (0.8 * 0.3 * 0.1 * 0.6 * 0.25 * 0.3 * 0.1)
mal 2, da zwei Lesarten

6 Statistik

Das Wort 'love' als Verb hat in WordNet 4 Lesarten. Ihre Hypothese ist, dass alle die gleiche Wahrscheinlichkeit haben. Sie machen eine Stichprobe von 100 Vorkommen von 'love' als Verb und bekommen folgende Ergebnisse: 10,20,30,40 (Häufigkeit für Lesart). Wie können Sie Ihre Hypothese testen? Welche Verteilung verwenden Sie? Formulieren Sie eine Nullhypothese und eine statistische Prüfgrösse (sie muss nicht ausgerechnet werden, Angabe der instantiierten Formel genügt).
Chi-Quadrat: Erwartungswert jeweils 25
Prüfwert: $((10-25)^2 + (20-25)^2 + 30 - 25^2 + 40 - 25)^2)/25$