Klausur 2012: Quantitative Methoden der CL (70 Min., 60 Pkt.

1 Allgemeine Fragen (20 Min., 20 Pkt.)

1. Wie sieht die Bayes'sche Formel angewandt auf die statistische maschinelle Übersetzung aus?
2. Was ist mit dem Begriff Sprachmodell gemeint?
3. Was ist mit dem Begriff sparse data gemeint?
4. Erläutern Sie den Begriff der Verteilung.
5. Diskutieren Sie: $P(A \cap B) = P(A) * P(B)$

2 Tagging mit HMM (10 Min., 6 Pkt.)

Berechnen Sie die Wahrscheinlichkeit der Lesart 'pro aux verb' für den Satz 'I can can', i.e.

$$P(pro\ aux\ verb \mid I\ can\ can) = \prod_{i=1,n} P(t_i \mid t_{i-1}) \ * \ P(w_i \mid t_i)$$

Uebergangswahrscheinlichkeiten:

	pro	aux	inf	verb	det	noun	start
start	1	0	0	0	0	0	0
pro	0	3/4	0	1/4	0	0	0
aux	0	0	1/3	1/3	1/3	0	0
inf	0	0	0	0	1	0	0
verb	0	0	0	0	1	0	0
det	0	0	0	0	0	1	0
noun	0	0	0	0	0	0	0

Emissionswahrscheinlichkeiten:

	she	i	can	come	has	a	the	fish	will
pro	3/4	1/4	0	0	0	0	0	0	0
aux	0	0	1/3	0	1/3	0	0	0	1/3
inf	0	0	1/2	1/2	0	0	0	0	0
verb	0	0	1	0	0	0	0	0	0
det	0	0	0	0	0	1/3	2/3	0	0
noun	0	0	1/3	0	0	0	0	2/3	0

3 MLE, Smoothing (10 Min., 8 Pkt.)

Schätzen Sie die Wahrscheinlichkeiten der 26 Buchstaben, i.e. P(x) mit x in $\{a, b, c, ..., x, y, z\}$ anhand der folgenden Wörter: [der,drei,derbe]. Verwenden Sie Witten-Bell-Smoothing.

Witten-Bell:

- die Wahrscheinlichkeitsmasse für nicht-gesehene Ereignisse ist: $\frac{T}{N+T}$ mit T=Anzahl der Types und N=Anzahl der Token
- Diese müssen auf alle nicht-gesehenen Ereignisse verteilt werden: $\frac{T}{Z(N+T)}$ wobei Z die Menge der Nullereignisse ist
- alle *beobachteten* Ereignisse haben die Schätzung (c_i als Zählfunktion): $\frac{c_i}{N+T}$

4 Entscheidungsbaum (10 Min., 8 Pkt.)

Gegeben folgende Vektoren und ihre Klassenzugehörigkeit:
⟨2,j,i,j⟩ Klasse = 1
⟨2,g,2,g⟩ Klasse = 0
⟨b,j,2,g⟩ Klasse = 1
⟨i,g,i,j⟩ Klasse = 0
⟨i,j,2,j⟩ Klasse = 1

a) Wie gross ist die Entropie der Beispielmenge?
b) Wie gross ist der Entropiegewinn, wenn wir die 4. Dimension als Wurzelknoten vorsehen?
c) Bei welcher Dimension (Attribut) ist der Entropiegewinn maximal?

Formeln:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

$$Gain(S, A) \equiv Entropy(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

wobei S_v die Menge der Beispiele ist, die für Attribut A den Wert v aufweisen.

Es reicht, wenn Sie die Formel instantiieren, Sie brauchen den Wert nicht ausrechnen.

5 Grenzwerttheorem (10 Min., 8 Pkt.)

Kann man den Mittelwert einer Stichprobe bedenkenlos als Schätzung für den Mittelwert der Grundgesamtheit nehmen? Diskutieren Sie die Problematik.

6 Chi-Quadrat (10 Min., 10 Pkt.)

Sind die verschiedenen Verb-Partikel-Konstruktionen im Englischen (vgl. die Beispielsätze) zufällig oder gibt es Präferenzen für eine der beiden.

Beispiele:
I: He picked up the book
II: He picked the book up

Folgende Häufigkeiten:
I: 247
II: 150

H_0 sei: Die Häufigkeiten der Konstruktionen sind identisch, die Variation in der Stichprobe ist zufällig.

a) Testen Sie anhand Chi-Quadrat. Instantiieren Sie die Formel.

$$\chi^2_{(k-1)} = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j}$$

wobei:
e_j = Erwartungswert
f_j = Häufigkeit
k = Anzahl der unabhäng. Versuche

b) Gegeben folgende Tabelle (df steht für Freiheitsgrad) und nehmen wir an, der Prüfwert sei 6.2 (was er definitiv nicht ist). Könnten wir dann die Nullhypothese verwerfen und falls ja, auf welchem Signifikanzniveau? Geben Sie unbedingt an, welche Zeile Sie zugrundegelegt haben.

		p=0.05	p=0.01
1	df=1	3.841	6.635
2	df=2	5.991	9.21
3	df=3	7.815	11.345