

Übungsblatt 6

10. Dezember 2008

Hinweis: Soweit nicht anders angegeben, gibt es für jede korrekt bearbeitete Teilaufgabe einen Punkt. Die Abgabe der Hausübungen ist bis spätestens zum Beginn der nächsten Vorlesung möglich – entweder persönlich direkt vor der Vorlesung oder per Einwurf in den Briefkasten des Instituts (Informatikzentrum, zweiter Stock, vor Raum 238).

Aufgabe 11 (Language Models I)

In dieser Aufgabe betrachten wir das folgende 2-Wort-Bigramm-Modell:

$$P(A) = 0,7 \quad P(B) = 0,3 \quad P(A|A) = 0 \quad P(B|A) = 1 \quad P(A|B) = 0.8 \quad P(B|B) = 0.2$$

Welche Wortfolgen der Länge 4 können mit diesem Modell erzeugt werden? Mit welcher Wahrscheinlichkeit geschieht dies jeweils?

Aufgabe 12 (Language Models II)

In dieser Aufgabe betrachten wir die folgende Dokumentensammlung über dem Vokabular {click, go, the, shears, boys, metal, here}:

Dokumentnummer	Inhalt
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

Schätzen Sie für jedes Dokument ein Unigramm-Modell und beantworten Sie damit die Anfragen „click“ und „click shears“. Verwenden Sie zur Schätzung der Modellparameter ...

- ... den Maximum-Likelihood-Schätzer. (2 Punkte)
- ... den Maximum-Likelihood-Schätzer mit einfacher Glättung (wählen Sie $\alpha = 0,5$). (2 Punkte)
- ... den Maximum-Likelihood-Schätzer mit linearer Glättung (wählen Sie $\lambda = 0.3$). (2 Punkte)

Aufgabe 13 (Retrievalbewertung)

- a) Grenzen Sie die Begriffe Pertinenz und Relevanz voneinander ab. Verwenden Sie zur Veranschaulichung ein selbstgewähltes Beispiel (also nicht das Fahrplan-Beispiel aus der Vorlesung).
- b) In der Vorlesung wurde ein Informationsbedürfnis aus der TREC-Ad-Hoc-Kollektion vorgestellt („Endangered Species (Mammals)“, Folie 56). Formulieren Sie auf ähnliche Weise ein selbstgewähltes Informationsbedürfnis zu einem aktuellen Thema aus den Nachrichten. Verwenden Sie dazu denselben Aufbau wie in der TREC-Kollektion (title, description, narrative). (2 Punkte)