Einführung in die Computerlinguistik **WS 2005/2006**
M. Pinkal, S. Pado

Übungsblatt 5 (Abgabe 31.01.2006)

1. **Evaluation (3P)**
   Die Folien zur Vorlesung zeigen eine Beispielevaluation für eine binäre Klassifikationsaufgabe (Folie 35ff.), konzentrieren sich aber auf die Klasse Adj.
   a. Berechnen Sie Precision, Recall und F-Score für die Klasse NAdj.
   b. Vergleichen Sie die Ergebnisse für NAdj mit den Ergebnissen für Adj. Wie gut ist der Klassifikator Ihrer Meinung nach tatsächlich, um Adjektive zu erkennen?

2. **Statistische Modellierung (4P)**
   Herr P. trainiert den statistischen Klassifikator aus den Vorlesungsfolien auf einigen hundert Sätzen aus dem TIGER-Korpus und erhält folgende Frequenztabelle:

Selbst Artikel?	Nächstes Wort kapitalisiert?		Adj	NAdj
Falsch	Falsch		66	10533
Falsch	Wahr		787	4966
Wahr	Falsch		0	565
Wahr	Wahr		0	1162

   a. Wie viele Klassen gibt es, wie heißen sie?
   b. Wie viele Features gibt es? Wie viele Werte haben die Features jeweils? Wie groß ist der Ereignisraum?
   c. Für welchen Anteil des gesamten Ereignisraumes hat der Klassifikator Trainingsinstanzen gesehen? Was bedeutet das für die Abdeckung des Klassifikators auf neuen Daten?
   d. Paraphrasieren Sie die Bedeutung des Modells als Menge von Regeln, die den einzelnen Ereignissen Klassen zuordnen. Wie denken Sie über die Nützlichkeit des Klassifikators?
   e. Welches Grundproblem hat also die naive statistische Klassifikation, wenn manche der Klassen sehr viel kleiner sind als andere Klassen?

3. **Probabilistische kontextfreie Grammatiken (4P)**
   a. Was kann eine probabilistische Grammatik beschreiben, das eine symbolische nicht kann?
   b. Wie liest man aus einer Baumbank Regelwahrscheinlichkeiten ab?
   c. Nennen Sie einen linguistischen Zusammenhang, der von Regeln der Form X-> Y Z und den dazugehörigen Wahrscheinlichkeiten nicht modelliert werden kann und erklären Sie, wieso nicht. Wie kann man das Regelformat erweitern, um den Zusammenhang zu modellieren? Wie berechnet man die Wahrscheinlichkeit für das neue Regelformat? Was für ein Problem tritt bei dem neuen Regelformat auf bzw. verstärkt sich?