Schriftliche Prüfung zur Übung
Statistische Methoden in der maschinellen Sprachverarbeitung
WS 2014/15
Helmut Schmid

1. Wortart-Tagger mit Bigramm-HMM

Implementieren Sie (in Python, Perl oder Java) einen Wortart-Tagger auf Basis eines Bigramm-Hidden-Markow-Modelles. Sie sollen eine Funktion **viterbi(tokens)** implementieren, welche die beste Tagfolge eines Satzes berechnet und als Liste zurückgibt. Das Argument der Funktion (tokens) ist eine Liste von Tokens.

- **contextProb(tag1, tag2)** gibt die Kontextwahrscheinlichkeit p(tag2|tag1) zurück. Das Satzgrenzentag soll $\langle s \rangle$ sein.
- **lookup(word)** gibt eine Liste von Tag-Wahrscheinlichkeit-Paaren zurück, wobei es sich bei der Wahrscheinlichkeit um $p(word|tag)$ handelt.

Ihr Programm soll die beste Tagfolge ausgeben. Sie müssen nicht mit Logarithmen arbeiten. (15 Punkte)

2. Wortbedeutungsdesambiguierung

Implementieren Sie einen Wortbedeutungsdesambiguierer auf Basis eines Naive-Bayes-Modelles. Auch hier schreiben Sie eine Funktion **desamb(word, tokens)**, welche das zu desambiguierende Wort und einen Text in Form einer Liste von Tokens als Argumente nimmt. Folgende Funktionen sind gegeben:

- Die Funktion **senses(word)** liefert die Liste der Bedeutungen des in "word" gespeicherten Wortes. (Bspw. könnte für das Wort *Bank* die Liste {*Bank1, Bank2, Bank3*} zurückgegeben werden.
- **wordProb(word, sense)** liefert die Wahrscheinlichkeit des Wortes "word" im Kontext der Bedeutung "sense" des zu desambiguierenden Wortes. Der Rückgabewert der Funktion ist immer größer als Null.
- **priorProb(sense)** liefert die Apriori-Wahrscheinlichkeit der Bedeutung "sense".

Sie sollen in der Tokenliste nach Vorkommen des zu desambiguierenden Wortes suchen und diese auf Basis der Kontextwörter mit maximalem Abstand 50 desambiguieren. Ausgegeben werden soll jeweils die Position des desambiguierten Wortes und die zugewiesene Bedeutung. Sie können die Ergebnisse entweder auf die Konsole ausgeben oder in einer Liste von Paaren (Position, Bedeutung) zurückgeben.

(15 Punkte)

Viel Erfolg!