Sprachtechnologie

2. Übung

Prof. Tobias Scheffer
Thomas Vanck

Sommer 2010

Ausgabe am: 3.05.10
Besprechung am: 10.05.10

Aufgabe 1 *N-Gramme*

Gegeben sei das folgende Trainingskorpus: *„Ich habe Hunger und Durst. Aber ich habe kein Essen und kein Getränk dabei. Ausserdem habe ich kein Geld. Ich muss zuerst Geld besorgen.“* und das Testkorpus: *„Ich habe kein Geld.“*.

- Bestimmen Sie die Vokabulargröße k des Trainingskorpus.
- Bestimmen Sie den ML- und MAP-Schätzer für die Eingramm-$P(w_t)$ und Zweigrammwahrscheinlichkeiten $P(w_t|w_{t-1})$ unter der Annahme, dass alle Parameter des Dirichlet-verteilten Priors $\alpha = 2$ sind mit Hilfe des Trainingskorpus.
- Geben Sie mit Hilfe der jeweils geschätzten Parameter die Gesamtwahrscheinlichkeit für das Testkorpus an.

Aufgabe 2 *Autovervollständigung*

Autovervollständigungssysteme wie sie beispielsweise bei Texteditoren Anwendung finden, sollen dem Benutzer Wörter vorschlagen bevor sie vollständig eingeben wurden sind. Angenommen Sie haben die Aufgabe eine derartig intelligente Eingabehilfe zu entwickeln. Diskutieren Sie Lösungsansätze für diese Aufgabenstellung. Zeigen Sie dabei, wie Sie aus einer Eingabe $w_1, \ldots, w_T$ (die einzelnen w_i sind hier Buchstaben) die wahrscheinlichste Fortsetzung $argmax_{w_{T+1},\ldots,w_{t+k}} P(w_{T+1}, \ldots, w_{t+k}|w_T, \ldots, w_1)$ berechnen könnten.

Aufgabe 3 *T9*

Die Eingabehilfe T9 ermöglicht die Texteingabe auf Zifferntastaturen. Diskutieren Sie, wie Sie T9 mit Hilfe eines N-Gramm-Modells auf Buchstabenebene implementieren können. Das System soll stets die wahrscheinlichste Wort für die getippten Ziffern ermitteln. Wie rechenaufwändig ist Ihre Lösung?