

Information Retrieval

Sascha Kriewel

sascha.kriewel@uni-duisburg.de

Übungsblatt 7

Bearbeitung bis **13. Juni 2005**
keine Abgabe

Aufgabe 1: Part of Speech Tagging

Es gibt eine Reihe von Taggern, die im Internet ausprobiert werden können. Die folgenden Links stellen eine Auswahl dar. Probiere zumindest drei der Tagger aus der Liste aus und vergleiche sie. Nimm zum Testen jeweils den gleichen Text (deutsch und/oder englisch). Worin unterscheiden sich die Ergebnisse? Versuche auch etwas über die Arbeitsweisen der Tagger herauszufinden.

- <http://www.ifi.unizh.ch/CL/tagger/>
- http://www.ling.gu.se/~lager/Home/brilltagger_ui.html
- <http://www.connexor.com/demos.html>
- <http://www.coli.uni-sb.de/~thorsten/tnt>
- <http://ilk.kub.nl/~zavrel/tagtest.html>
- <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

Aufgabe 2: Hidden Markov Models

Die Wahrscheinlichkeit, dass ein Wort mit dem Tag t^k auf ein Wort mit dem Tag t^j folgt, sei beschrieben durch

$$P(t^k|t^j) = \frac{C(t^j, t^k)}{C(t^j)}$$

Die Wahrscheinlichkeit, dass ein Wort mit dem Tag t^j ein bestimmtes Wort w emittiert, sei beschrieben durch

$$P(w|t^j) = \frac{C(t^j, w)}{C(t^j)}$$

Die Abschätzung der *Maximum Likelihood* für eine bestimmte Zuordnung von Tags zu einem Satz $w^1 \dots w^n$ ergibt sich dann als:

$$\prod_{i=1}^n P(w^i|t^i)P(t^i|t^{i-1})$$

Dabei gelten die folgenden Konventionen:

- $C(t^j)$ Anzahl Vorkommen des Tags t^j im Trainingskorpus
- $C(t^j, t^k)$ Anzahl Vorkommen des Tags t^j gefolgt von t^k im Trainingskorpus
- $C(t^j, w)$ Anzahl Vorkommen des Wortes w getaggt als t^k

Nach dem Prozessieren eines englischsprachigen Trainingskorpus seien die folgenden Häufigkeiten gegeben. Dabei sind: AT – Artikel, BEZ – das Wort *is*, IN – Präposition, NN – Nomen, VB – Verb, STZ – Satztrennzeichen (.,!?:;).

Tabelle 1: Vorgängertag (Zeile) → Nachfolgertag (Spalte)

	AT	BEZ	IN	NN	VB	STZ	Summe
AT	0	0	0	35000	0	15	35015
BEZ	1700	0	370	140	0	50	2260
IN	31000	0	1200	15000	0	170	47370
NN	1000	3500	37000	10000	550	17000	59050
VB	5000	35	4000	1200	100	1250	11585
STZ	7000	70	3930	1000	800	0	12800

Tabelle 2: Wort (Zeile) → Tag (Spalte)

	AT	BEZ	IN	NN	VB	STZ	Summe
master	0	0	0	400	250	0	650
is	0	9000	0	0	0	0	9000
move	0	0	0	25	125	0	150
on	0	0	5000	0	0	0	5000
the	60000	0	0	0	0	0	60000
.	0	0	0	0	0	40000	40000

Bestimme anhand der gegebenen Parameter die Wahrscheinlichkeiten für die folgenden Tagzuordnungen zu dem Satz „*The master put chairs on the table*“. Überlege wie dabei die im Lernkorpus nicht vertretenen Wörter behandelt werden sollen.

- AT – NN – VB – NN – IN – AT – NN
- AT – VB – NN – VB – IN – AT – NN

Aufgabe 3: Informationsextraktion

Nimm an, Du möchtest Informationsextraktion auf Beschreibungen von Kinofilmen betreiben. Grundlage könnten Rezensionen von Filmkritikern, Werbematerial der Verreiber oder Zusammenfassungen in Kinoprogrammen sein. Suche online nach einer geeigneten Quelle.

Überlege Dir dann ein Schema oder Template für die zu extrahierende Information. Welche Daten lassen sich extrahieren?

Führe dann beispielhaft die Schritte der Informationsextraktion auf einer Filmbeschreibung aus Deiner Quelle durch. Für das *PoS-Tagging* kannst Du einen der Tagger aus Aufgabe 1 heranziehen. Finde Koreferenzen, fülle die Schemata auf und führe die Ergebnisse zusammen.