

Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

Sommersemester 2010, Prof. Dr. Udo Hahn, Erik Fäßler

Übungsblatt 5 vom 02.06.2010

Abgabe bis 08.06.2010, 23.59 Uhr; per Email (Standard Dateiformat: ps, pdf, oder doc) an erik.faessler@uni-jena.de

Aufgabe 1 : Annotation

Betrachten Sie folgenden Text:

“Die letzten Sekunden des Bundespräsidenten Horst Köhler lösen widersprüchliche Gefühle aus. Wie er da Hand in Hand mit seiner Frau dem Ausgang des Schlosses Bellevue zustrebte, das hatte etwas Anrührendes. Man fragte sich, wie übel dem Mann mitgespielt worden sein muss, dass er so geht.”

Die Wörter des oben gegebenen Texts soll auf ihre Wortarten hin annotiert werden. Als mögliche Wortarten dient das folgende Tag-Set (eine Menge von Tags oder Labeln):

Tag	Beschreibung	Beispiel
ADJA	attributives Adjektiv	[das] große [Haus]
ADJD	adverbiales oder prädikatives Adjektiv	[er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
K	Konjunktion	weil, dass, damit, wenn, ob
N	Nomen	Tisch, Herr, Hans
P	Pronomen	er, mein, sich, jener, kein [Mensch],irgendein [Glas]
PR	Präposition	in [der Stadt], ohne [mich]
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
V	Verb	gehen, stand, gelegen
\$,	Komma	,
\$.	Satzbeendende Interpunktion	. ? ! ; :

Um jedem Wort eine Wortart zuzuweisen, bauen Sie eine Tabelle auf, in der die jeweilige Zuweisung ersichtlich ist. Um Platz zu sparen, können Sie mehrere Spalten nebeneinander abbilden. Fügen Sie außerdem eine Spalte hinzu, um Chunks zu annotieren. Verwenden Sie für die Chunk-Annotation das in Vorlesung und Übung gezeigte IOB (BIO) Format.

Aufgabe 2 : NP-Parsing

Sehen Sie sich den ersten Satz des Texts aus der ersten Aufgabe an. Sie haben bereits Chunks im IOB-Format identifiziert. Schreiben Sie eine formale Grammatik, um die Nominalphrasen dieses Satzes syntaktisch zu analysieren und führen sie die Analyse durch.

Aufgabe 3 : Automatische Wahrscheinlichkeitsschätzung

1. Schreiben Sie in Pseudocode einen Algorithmus, der den annotierten Corpus der ersten Aufgabe entgegen nimmt und die absoluten sowie relativen Häufigkeiten jedes Types (also der unterschiedlichen Wörter) berechnet.
2. Sei V die Menge aller Types aus dem Text aus Aufgabe 1 (also das Vokabular). Schreiben Sie ein Programm im Pseudocode, das für jedes Wort $w_i \in V$ die bedingte Wahrscheinlichkeit berechnet beobachtet zu werden, gegeben eines unmittelbar zuvor vorkommenden Wortes $w_j \in V$, d.h. $P(w_i|w_j)$, wobei $i, j \in \{1, \dots, |V|\}$.