

Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

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

Übungsblatt 6 vom 10.06.2010

Abgabe bis 15.06.2010, 23.59 Uhr; per Email (Standard Dateiformat: ps, pdf, oder doc) an erik.faessler@uni-jena.de

Aufgabe 1 : Chunking

Gegeben sei folgenden Text:

Eine unpassendere Trophäe hätte er wohl nicht bekommen können. Ausgerechnet einen spitzen Siliziumkristall bekam Michael Grätzel am Mittwochabend von der finnischen Staatspräsidentin Tarja Halonen überreicht. Die Skulptur soll den Gipfel der Technik repräsentieren und ist das Symbol des Millennium-Technologiepreises, mit insgesamt 1,1 Millionen Euro einer der höchst dotierten Technikpreise der Welt.

1. Die Wörter des oben gegebenen Texts sollen analog des letzten Übungsblattes auf ihre Wortarten hin annotiert werden. Als mögliche Wortarten dient das folgende Tag-Set:

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
PART	Partikel	nicht, aus
KARD	Kardinalität	zehn, 100
\$,	Komma	,
\$.	Satzbeendende Interpunktion	. ? ! ; :

2. Führen Sie nun NP-Chunking auf dem PoS-Annotierten Text durch. Annotieren sie dabei nur Basis-NP-Chunks, also einfache, ungeschachtelte Nominalphrasen. Verwenden Sie das IOB-Format.
3. Schreiben Sie nun Regeln in Regulären Ausdrücken, um das Chunking automatisch durchzuführen. Orientieren Sie sich dabei an Folie 67 des fünften Teils der Vorlesung. Es ist Ihnen auch freigestellt, das Format zu verwenden, das auf <http://nltk.googlecode.com/svn/trunk/doc/howto/chunk.html> beschrieben ist. Beantworten Sie dazu folgende Fragen:
 - Was ist der Unterschied zwischen den beiden Darstellungen?
 - Welche der Darstellungen benötigt einen Konvertierungsschritt, bevor tatsächlich ein regulärer Ausdruck entsteht, der zur Erkennung von Chunks verwendet werden kann?
 - Beschreiben Sie kurz, was bei dieser Konvertierung geschieht.
4. Geben Sie nun Reguläre Ausdrücke an, um ein Chinking auf dem obigen Text durchzuführen, so dass die Basis-NP-Chunks übrig bleiben.