

Probeklausur

Aufgabe 1

- (a) Lexikalische Information in Sprachverarbeitungssystemen kann alternativ in Wortformenlexika und in Stammlexika bereitgestellt werden. Wieso ist es für das Deutsche wichtiger als für das Englische, mit einem Stammlexikon zu arbeiten?
- (b) Die Kontinuität und die Varianz des Schallsignals sind Probleme für die Spracherkennung. Geben Sie für Kontinuität und Varianz jeweils zwei unterschiedliche Aspekte an.
- (c) Wann ist Kontextwissen bei der Rechtschreibkorrektur nötig? Erläutern Sie an zwei Beispielen.
- (d) Nennen Sie drei potentielle Probleme, die ein maschinelles Übersetzungsprogramm (z.B. ins Englische) mit dem folgenden Satz haben kann: *Das Mädchen las seiner Mutter auf der Bank vor dem Haus aus dem Buch vor.*
- (e) Setzen Sie die folgenden Begriffe durch die passenden semantischen Relationen miteinander in Beziehung: *Musikinstrument, Geige, Saite, Streichinstrument, Bratsche, Trompete.*
- (f) Geben Sie eine Analyse mit Hilfe des X-Bar-Schemas für die NP *der große dicke Baum hinter dem Haus* an. Denken Sie an die Kantenbeschriftungen.
- (g) Geben Sie zwei Zerlegungen für das Wort *Abteilungen* an, die ein Morphologiesystem erkennen könnte. Benennen Sie die einzelnen Morpheme (Stamm, Flexionsendung,...).
- (h) Kommentieren Sie, welche besonderen Schwierigkeiten der folgende Satz für ein TTS-System macht. Nennen Sie 3 Punkte. *Bei der Bundestagswahl am 27.9.2009 erhielt die CDU im Wahlkreis Coesfeld ca. 48% der Zweitstimmen.*
- (i) Nennen Sie jeweils einen Vorteil des Transfer-Modells gegenüber dem Interlingua-Modell in der maschinellen Übersetzung und umgekehrt.

Aufgabe 2

In folgenden Beispielen sind abgetrennte Verbzusätze (VZ) hervorgehoben:

Det N V VZ	Der Zug fährt ab
PN V Det N VZ	Peter sammelt die Teller ein
Det N V Det N Det N VZ	Der Trainer teilt den Spielern die Positionen zu

Entwerfen Sie einen endlichen Automaten, der einfache Sätze mit abgetrennten Verbzusätzen erkennt. Akzeptiert werden sollen Folgen der Art: Nominalausdruck - Verb -

beliebig viele Nominalausdrücke – Verbzusatz. Nominalausdrücke sollen immer aus einem Eigennamen oder einem Artikel gefolgt von einem Nomen bestehen. Geben Sie einen zweiten Automaten an, der zwischen Verb und Verbzusatz maximal zwei Nominalausdrücke akzeptiert, sich ansonsten aber wie der erste Automat verhält.

Aufgabe 3

- (a) Gegeben sind die beiden Automaten $A_1 = \langle K_1, \Sigma, \Delta_1, s_1, F_1 \rangle$ und $A_2 = \langle K_2, \Sigma, \Delta_2, s_2, F_2 \rangle$ mit

$$K_1 = \{1, 2\}, K_2 = \{3, 4\}$$

$$\Sigma = \{a, b\}$$

$$s_1 = 1, s_2 = 3$$

$$F_1 = \{2\}, F_2 = \{4\}$$

$$\Delta_1 = \{\langle 1, b, 1 \rangle, \langle 1, a, 2 \rangle, \langle 2, b, 2 \rangle\}, \Delta_2 = \{\langle 3, b, 3 \rangle, \langle 3, aa, 4 \rangle, \langle 4, b, 4 \rangle\}$$

Beschreiben Sie jeweils in Worten die Sprachen L_1 und L_2 , die die Automaten erkennen.

- (b) Konstruieren Sie den Vereinigungsautomaten A_{12} , der die Vereinigung von L_1 und L_2 beschreibt. Es reicht den Automaten als Graphen darzustellen, Sie müssen ihn nicht formal spezifizieren.
- (c) Determinisieren Sie A_{12} nach dem in der Vorlesung beschriebenen dreistufigen Verfahren.

Aufgabe 4

Gegeben sei eine kontextfreie Grammatik mit den folgenden Produktionsregeln:

$$\begin{array}{ll} S \rightarrow NP VP & V \rightarrow \text{spielte, gähnte} \\ VP \rightarrow V NP & N \rightarrow \text{Musiker, Sinfonie, Paukenschlag, Klavier} \\ VP \rightarrow V & Prp \rightarrow \text{auf, mit} \\ VP \rightarrow VP PP & Det \rightarrow \text{der, dem, den, die, der} \\ NP \rightarrow NP PP & \\ PP \rightarrow Prp NP & \\ NP \rightarrow Det N & \end{array}$$

- (a) Leiten Sie den Satz *Der Musiker spielte die Sinfonie mit dem Paukenschlag.* auf zwei unterschiedliche Weisen ab. Geben Sie dazu die beiden Strukturbäume an, die aus der Ableitung resultieren.
- (b) Kann ein Top-Down Parser, der die Regeln in der angegebenen Reihenfolge ausprobieren, Probleme mit Linksrekursion bekommen? Begründen Sie.

Aufgabe 5

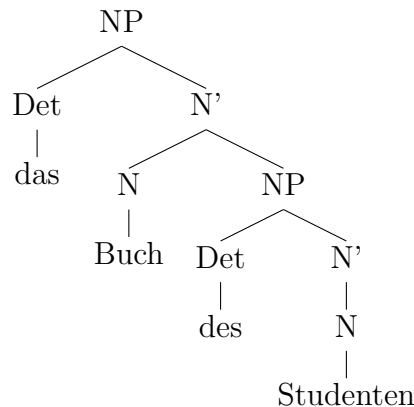
Stellen Sie durch Angabe der einzelnen (Teil-)Bäume und des zugehörigen Stacks dar, wie ein Top-Down-Parser den Satz *Der Musiker spielte* in der kürzestmöglichen Weise abarbeitet (das heißt, jeweils zielführende Regeln der Grammatik in Aufgabe 6 anwendet).

Aufgabe 6

(a) Die folgenden Regeln beschreiben NPs mit einem Genitivattribut (z.B. „*das Buch des Studenten*“).

$$\text{NP} \rightarrow \text{Det N}'$$

$$\text{N}' \rightarrow \text{N NP}$$

$$\text{N}' \rightarrow \text{N}$$


Diese Regeln generieren über. Erläutern Sie dies durch Angabe von 2 (grammatisch inkorrekten) Beispielen, wobei der Fehler jeweils auf unterschiedliche Art und Weise zustandekommen soll.

(b) Ergänzen Sie geeignete Merkmalsconstraints für die Regeln als Gleichungen, so dass nur noch grammatisch korrekte NPs erkannt werden.

Aufgabe 7

Beschreiben Sie die Funktionen von Predictor und Scanner im Earley-Algorithmus kurz und informell. Geben Sie jeweils ein Beispiel für die Funktionsweise.

Aufgabe 8

Die folgende Tabelle zeigt die Evaluation eines Klassifikators für die Erkennung von Fehlern in einem Rechtschreibkorrektursystem.

	echter Fehler	echtes korrektes Wort
klass. als Fehler	40	60
klass. als korrektes Wort	10	390

- (a) Der F-Score für die Klasse „korrektes Wort“ beträgt ca. 0,92. Bedeutet das, dass der Klassifikator besonders gut ist? Begründen Sie.
- (b) Berechnen Sie den F-Score für die Klasse Fehler.
- (c) Halten Sie es bei diesem System für wichtiger für die Klasse *Fehler* auf Präzision oder auf Recall zu optimieren? Begründen Sie.

Aufgabe 9

Mit Hilfe von Kontextwörtern soll ein Klassifikator zur Word-Sense-Disambiguierung für die beiden Lesarten von *Ball* gelernt werden. Das Trainingskorpus besteht aus 50 Dokumenten für $Ball_1$ und 50 Dokumenten für $Ball_2$.

Dabei ermittelt man die folgenden Kontextwortfrequenzen und beobachtet danach ein Auftreten von *Ball* in dem Satz „Nach dem ersten Tanz warf sie Ihre Schuhe in die Ecke und verließ den Ball“.

	Ball ₁	Ball ₂		???
Tor	33	2		
werfen	25	5		
Tanz	7	46		
Schuh	12	10		

Außerdem gelten a-priori-Wahrscheinlichkeiten von $P(Ball_1) = 0,9$ und $P(Ball_2) = 0,1$. Ergänzen Sie das Merkmalsmuster für das beobachtete Auftreten von *Ball* und berechnen Sie die wahrscheinlichere Lesart. Nehmen Sie an, dass ein Stemmer verwendet wird.

Sie müssen die Probeklausur nicht abgeben. Sie wird am nächsten Dienstag in der Vorlesung besprochen.