

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) Nennen Sie jeweils einen Vorteil des Transfer-Modells gegenüber dem Interlingua-Modell in der maschinellen Übersetzung und umgekehrt.
- (f) Setzen Sie die folgenden Begriffe durch die passenden semantischen Relationen miteinander in Beziehung: *Musikinstrument, Geige, Saite, Streichinstrument, Bratsche, Trompete.*
- (g) 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. Geben Sie auch die Dependenzanalyse an.
- (h) Geben Sie zwei Zerlegungen für das Wort *Abteilungen* an, die ein Morphologiesystem erkennen könnte. Benennen Sie die einzelnen Morpheme (Stamm, Flexionsendung,...).
- (i) 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 Duisburg ca. 48% der Zweitstimmen.*
- (j) Nennen Sie drei Qualitätskriterien für Dialogsysteme und erklären Sie an einem konkreten Beispiel (wie z.B. dem „Sprechenden Aufzug“) an, wie Sie diese Kriterien in einem Experiment mit menschlichen Testpersonen überprüfen würden.

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

$$\begin{aligned} K_1 &= \{1, 2\}, K_2 = \{3, 4\} \\ \Sigma &= \{a, b\} \\ s_1 &= 1, s_2 = 3 \\ F_1 &= \{2\}, F_2 = \{4\} \\ \Delta_1 &= \{\langle 1, b, 1 \rangle, \langle 1, a, 2 \rangle, \langle 2, b, 2 \rangle\}, \Delta_2 = \{\langle 3, b, 3 \rangle, \langle 3, aa, 4 \rangle, \langle 4, b, 4 \rangle\} \end{aligned}$$

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, nach dem besprochenen Verfahren. 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:

$S \rightarrow$	NP VP	$V \rightarrow$	spielte, gähnte
$VP \rightarrow$	V NP	$N \rightarrow$	Musiker, Sinfonie, Paukenschlag, Klavier
$VP \rightarrow$	V	$Prp \rightarrow$	auf, mit
$VP \rightarrow$	VP PP	$Det \rightarrow$	der, dem, den, die, der
$NP \rightarrow$	NP PP		
$PP \rightarrow$	Prp NP		
$NP \rightarrow$	Det N		

- (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 ausprobiert, Probleme mit Linksrekursion bekommen? Begründen Sie.

Aufgabe 5

- (a) 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 4 anwendet).
- (b) Wie wurde ein Bottom-up-Parser den Satz parsen? Geben Sie die einzelnen Stacks in der richtigen Reihenfolge an. Nehmen Sie an, dass Sie wann immer möglich zuerst reduzieren und dann shiften. In dieser Teilaufgabe können Sie auf Angabe der Teilbäume und benutzten Regeln verzichten.

Aufgabe 6

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

Aufgabe 7

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 8

- a) Was versteht man unter einem Perzeptron?

- b) Spezifizieren Sie ein neuronales Netz, mit drei Input- und einem Outputknoten (und beliebigen zusätzlichen Knoten). Die Inputknoten können genau die Werte 0 und 1 annehmen. Der Outputknoten soll genau dann feuern, wenn entweder alle drei oder genau der erste Outputknoten den Wert 1 hat.

Aufgabe 9

Sie arbeiten auf der Kommandozeile mit einer Version des Tigerkorpus, bei der in jeder Zeile ein Wort steht (*tiger.txt*), sowie mit einer POS-getaggten Version des Tiger-Corpus (*tiger_pos.txt*), bei der in jeder Zeile tab-separiert ein Wort und sein zugehöriger POS-Tag stehen. Außerdem haben Sie mit *tiger.t2c* ein für TGrep2 zugängliche Version des Tigerkorpus zur Verfügung.

- a) Wonach sucht der folgende Befehl?

```
cat tiger.txt | grep -e "^Haus.\+"
```

Nennen Sie zwei Beispielwörter, die der Befehl findet, falls das Wort tatsächlich in Tiger vorkommt.

- b) Geben Sie einen Befehl an, der Ihnen als Zahl ausgibt, wie oft die Wortform *Saarlandes* im Tigerkorpus vorkommt.
- c) Was macht der folgende Befehl?

```
cat tiger_pos.txt | cut -f2 | sort | uniq -c | sort
```

Beschreiben sie, was das allgemeine Ziel des Befehls ist, und was die einzelnen Schritte bewirken.

- d) Welche Art von Beispielen liefert Ihnen der folgende Befehl?

```
tgrep2 -c tiger.t2c '/^NP/ <1 /^ART/'
```

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