WS 2016/17

ECL: Einführung in die Computerlinguistik

Klausurbogen

Zeit: 90 Minuten

Beantworten Sie 4 von 5 Fragen.

- **Keine geschriebenen oder gedruckten Hilfsmittel oder deren elektronische Äquivalente erlaubt.**
- Die Anzahl der Punkte für jede Frage ist in Klammern nach der Frage angegeben. Bitte berücksichtigen Sie dies bei der Zeiteinteilung.
- Verschwenden Sie keine große Zeit bei der Fragenauswahl. Frage 1 geht über Textklassifikation und enthält die Transferaufgabe, Frage 2 ist über REs, Automaten und Tokenisierung, Frage 3 über Parsing mit CYK, Frage 4 über Tagging mit Viterbi und Frage 5 behandelt Semantik und die Evaluation von IR.
- Beantworten Sie kurze Fragen auf dem Klausurbogen und lange Rechenfragen auf den zur Verfügung gestellten Extrabögen. Geben Sie sowohl Klausurbogen als auch Extrabögen ab.
- Bitte füllen Sie Ihren Namen, Matrikelnummer und gewählte Fragen in die untenstehende Tabelle vor Abgabe des Examens. Bitte beachten Sie, dass auf allen abzugebenden Extrablättern auch Name und Matrikelnummer stehen müssen.

Name	
Matrikelnummer	
Gewählte Fragen	Punkte

Frage 1

Textklassifikation (inkl. Transferfrage)

(a) Gegeben sei das folgende Trainings- und Testset

	docID	Dokument	Klasse = *Politik*
Trainingsset	1	Merkel Schulz CDU CDU	+
	2	CDU	+
	3	Merkel	+
	4	WM Fußball	-
	5	WM Merkel	-
Testset	6	Fußball Merkel Merkel	?

Schätzen Sie einen vollständigen **Bernouilli** Naive Bayes Klassifizierer auf Basis der Trainingsdaten. Welche Klasse weist ihr Klassifizierer dem Dokument aus dem Testset zu? Geben Sie alle Berechnungen als Teil der Lösung mit ab. Benutzen Sie Laplace-Smoothing. Antwort bitte auf Extrabogen.

[8 Punkte]

(b) **Transferfrage.** Sie sollen (englische oder deutsche) Posts auf Internetforen nach dem Alter der Autoren klassifizieren. Jedem Post wird einer Kategorie von *jung, mittel, alt* zugeordnet. (Wir interessieren uns nicht dafür, welchen Jahreszeitraum *jung*, *mittel* oder *alt* umfassen.)

(i) Sie trainieren einen Klassifizierer (wie Naive Bayes) mit n-grammen als Merkmalen. Diskutieren Sie, ob nur Wort-ngramme oder auch Punktuations-ngramme für die Altersklassifizierung von Forumposts verwendet werden sollten.

[4 Punkte]

(ii) Nennen Sie ein (1) Merkmal, das für Altersklassifikation nützlich sein könnte, aber nicht durch n-gramme erfasst wird. Wie würden Sie dieses Merkmal formalisieren?

[4 Punkte]

[Frage 1 gesamt: 16 Punkte]

Frage 2

Reguläre Ausdrücke, Automaten, Tokenisierung

(a) Ein(e) Student(in) schreibt ein Programm, das alle Worttypen in einem Korpus zählt und dann eine Tabelle mit den "Häufigkeiten von Häufigkeiten" ausgibt. Die Ausgabe ist in untenstehender Tabelle zu sehen. Man kann diese Tabelle lesen als *es gibt 41 Worttypen, die genau einmal auftauchen; 91, die zweimal auftauchen, 3 die 120 Mal auftauchen . . . .*

Worthäufigkeit	Häufigkeit der Häufigkeit
1	41
2	91
3	120
4	250
5	261
6	260
. . .	. . .
10	140
. . .	. . .
20	90
. . .	. . .
30	40

Erläutern Sie, warum das Program fehlerhaft sein muss. Welche Art der Verteilung würden Sie stattdessen erwarten? Formeln müssen nicht angegeben werden.

[4 Punkte]

(b) Konstruieren Sie für die folgenden Zeichenketten einen regulären Ausdruck, der in der Lage ist, sämtliche Ausdrücke der Sprache zu erkennen.

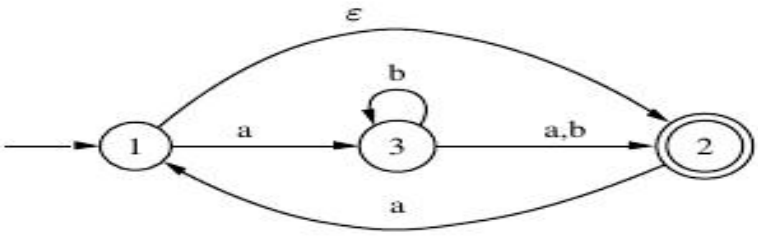
(i) Die Menge aller Zeichenfolgen über dem Alphabet {a,b,c}, die *nacheinander* beliebig viele oder kein *a*, dann genau ein *b* und beliebig viele oder kein *c* enthalten.

[1 Punkte]

(ii) Die Menge der Zeichenfolgen über dem Alphabet {a,b}, mit einer durch drei teilbaren Anzahl von b's. (Null ist auch durch drei teilbar).

[3 Punkte]

(c) Konvertieren Sie den untenstehenden nicht-deterministischen FSA in einen deterministischen mit der Powersetkonstruktion. Geben Sie die Schritte der Powersetkonstruktion bitte vollständig an. Antwort bitte auf Extrabogen.



[8 Punkte]

[Frage 2 gesamt: 16 Punkte]

Frage 3

Grammatik, Parsing

(a) Gegeben sei die folgende Grammatik G:

- S $\rightarrow$ NP VP
VP
- NP $\rightarrow$ ADJ NP
NP NP'
Professoren
Studenten
- NP' $\rightarrow$ CONJ NP
- CONJ $\rightarrow$ und
oder
- ADJ $\rightarrow$ neugierige
faule
- VP $\rightarrow$ lesen
schreiben

Betrachten Sie folgenden Satz S1: *neugierige Professoren und Studenten lesen*. Alle Antworten zu dieser Frage bitte auf Extrabogen.

(i) Geben Sie die beiden möglichen Lesarten von S1 nach der Grammatik G an (mittels gezeichneter Parsebäume oder der geklammerten Schreibweise).

[4 Punkte]

(ii) Wenden Sie den nicht-probabilistischen CYK-Algorithmus an, um auf Basis der Grammatik G den Satz S1 zu parsen.

[6 Punkte]

(iii) Zeigen Sie, wo in der Chart die Satzambiguität ausgedrückt wird und wie man daraus die beiden Parses herauslesen kann.

[2 Punkte]

(iv) Im Beispielfall würde die Erweiterung auf eine PCFG und den probabilistischen CYK den Satz dennoch nicht disambiguieren können. Dies kann man an den Bäumen und der Definition von PCFG sehen, ohne konkrete Wahrscheinlichkeiten zu verwenden. Erklären Sie, warum das so ist.

[4 Punkte]

[Frage 3 gesamt: 16 Punkte]

Frage 4

Tagging (mit kleinem Viterbi)

(a) HMM Tagging benutzt mehrere Unabhängigkeitsannahmen. Nennen Sie eine (1) davon und diskutieren Sie deren Korrektheit.

[4 Punkte]

(b) Gegeben sei folgender Satz S1:

- S1: *Trump deals are . . .* (im Deutschen *Trump(s) Geschäfte sind . . .*).

Tabelle 1 gibt die Emissionswahrscheinlichkeiten an. Beispiel: $p(Trump|NP0) = 0.02$

Tabelle 1: Emissionswahrscheinlichkeiten

	Trump	deals	are
Eigennamen NP0	0.02	0	0
Singularnomen NN1	0.00001	0	0
Pluralnomen NN2	0	0.0015	0
Verben (3sg) VVZ	0	0.01	0
Formen von be: VBE	0	0	0.71

Tabelle 2 gibt die Übergangswahrscheinlichkeiten zwischen Tags an. Das erste Tag ist in der Zeile. Beispiel: $p(NP0|NN2) = 0.003$

Tabelle 2: Übergangswahrscheinlichkeiten

	NP0	NN1	NN2	VVZ	VBE
$< s >$	0.067	0.037	0.022	0.0007	0.001
NP0	0.21	0.067	0.016	0.013	0.002
NN1	0.008	0.08	0.01	0.012	0.002
NN2	0.003	0.012	0.005	0.002	0.03
VVZ	0.02	0.04	0.02	0.0005	0.0006

Wir benutzen kein Smoothing.

(i) Welche möglichen Tagsequenzen gibt es für S1 laut der Emissionstabelle?

[2 Punkte]

(ii) Ermitteln sie die wahrscheinlichste POS-Tag Sequenz für Satz S1 durch Anwendung des Bigramm HMM-Modells, das durch die Tabellen spezifiziert wird. Benutzen Sie hierfür den Viterbi-Algorithmus und zeigen Sie Ihre Trellis (entweder als Grafik oder durch Spezifikation der Zellen). Antwort bitte auf Extrabogen.

Wenn man anstatt mit Viterbi die Wahrscheinlichkeiten für alle Lesarten separat berechnet, kann man bis zu 5 Punkte erreichen.

[10 Punkte]

[Frage 4 gesamt: 16 Punkte]

Frage 5

Semantik und Information Retrieval Evaluation

(a) Sie bekommen das folgende kleine Korpus, das aus 3 Sätzen besteht:

S1: Wir essen Nudeln und Tomaten.
S2: Wir essen Nudeln in Afrika.
S3: In der Heimat essen wir Tomaten und Nudeln.

Messen Sie die Ähnlichkeit des Wortpaares *Nudeln-Tomaten*, mithilfe eines vektorbasierten Ähnlichkeitsmodells auf dem obigen Korpus. Benutzen Sie alle fünf Inhaltswörter als Vektorkomponenten (Spalten), den ganzen Satz als Fenstergröße und einfache Kookkurrenz als Assoziationsmaß.

Formel für Kosinus zweier Vektoren: $cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$

	essen	Nudeln	Tomaten	Afrika	Heimat
Nudeln					
Tomaten					

[3 Punkte]

(b) Anstatt einfacher Ko-okkurrenz benutzen vektorbasierte Ähnlichkeitsmodelle oft Positive Pointwise Mutual Information. Definieren Sie das Maß, erklären Sie seine Eigenschaften sowie Vor- und Nachteile.

Hilfestellung Formel für PMI $pmi(x, y) = log\frac{p(x,y)}{p(x)\cdot p(y)}$.

[5 Punkte]

(c) Erklären Sie anhand der nachstehenden Definitionen des Wortes *Gericht* die Begriffe *Homonymie* und *Polysemie*.

- öffentliche Institution, die vom Staat mit der Rechtsprechung betraut ist, Verstöße gegen Gesetze bestraft und Streitigkeiten schlichtet
- als Mahlzeit zubereitete Speise
- Gebäude, in dem ein Gericht untergebracht ist

[4 Punkte]

(d) **Evaluation in IR.** Sie arbeiten auf einer Dokumentenkollektion mit genau 4 relevanten Dokumenten für Ihre Suchanfrage. Zwei IR Systeme geben jeweils 10 Dokumente zurück (mit Ranking). Hierbei steht R jeweils für ein wirklich relevantes Dokument, und N für ein nicht-relevantes.

System 1	System 2
R	N
N	**R**
R	N
R	N
N	N
N	N
N	N
N	**R**
N	**R**
N	**R**

Berechnen Sie Precision und Recall für die beiden Systeme. Wie aussagekräftig ist dies für den Systemvergleich? Geben Sie eine alternative Evaluationsmetrik, die besser sein könnte (Berechnung nicht nötig).

[4 Punkte]

[Frage 5 gesamt: 16 Punkte]