Sommersemester 2019 14.05.2019

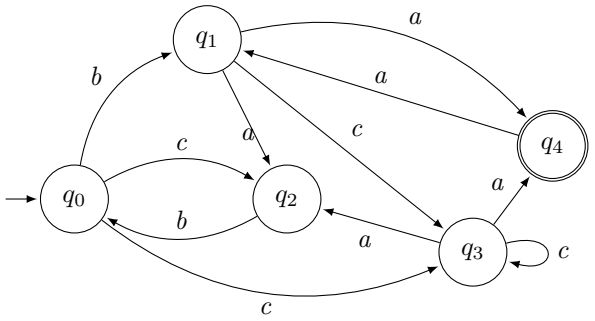
Einführung in die Computerlinguistik
Zwischenklausur

Laura Kallmeyer

Heinrich-Heine-Universität Düsseldorf

Erlaubte Hilfsmittel: Eine Din-A4 Seite mit Notizen.

Aufgabe 1 *(5 Pkte) Betrachten Sie den folgenden NFA:*



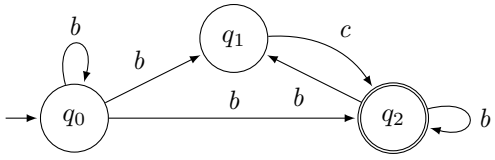
1. *Ist dieser Automat ein DFA, also deterministisch? Begründen Sie Ihre Antwort.*
2. *Welche der folgenden Wörter akzeptiert der Automat?*
   a) cbc *b) ccc* *c) cca*
3. *δ sei die Übergangsfunktion dieses Automaten. Geben Sie folgende Werte an:*
   (i) $\delta(q_0, c)$ *(ii) $\delta(q_1, a)$* *(iii) $\delta(q_4, c)$*
4. *$\hat{\delta}$ sei die reflexive transitive Hülle von δ wie in der Vorlesung definiert. Berechnen Sie die folgenden Werte:*
   (i) $\hat{\delta}(q_0, \varepsilon)$ *(ii) $\hat{\delta}(q_0, aa)$* *(ii) $\hat{\delta}(q_2, bcbcc)$*

Lösung:

1. Nein, da z.B. q_0 zwei ausgehende Kanten mit gleichem Label (nämlich c) hat. 1 Pkt
2. Nur c) wird akzeptiert. 1 Pkt
3. (i) $\delta(q_0, c) = \{q_2, q_3\}$ (ii) $\delta(q_1, a) = \{q_2, q_4\}$ (iii) $\delta(q_4, c) = \emptyset$ 1 Pkt
4. (i) $\hat{\delta}(q_0, \varepsilon) = \{q_0\}$ (ii) $\hat{\delta}(q_0, aa) = \emptyset$ (ii) $\hat{\delta}(q_2, bcbcc) = \{q_3\}$

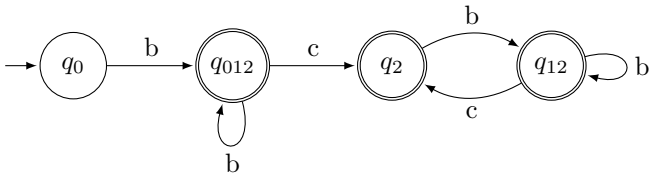
2 Pkte

Aufgabe 2 *(5 Pkte) Betrachten Sie nun den folgenden NFA:*



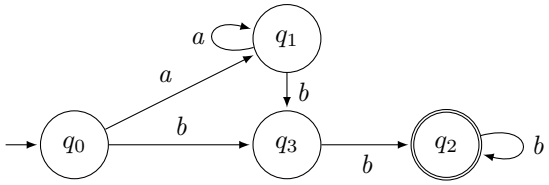
Konstruieren Sie einen äquivalenten DFA.

Lösung:



5 Pkte

Aufgabe 3 *(8 Pkte) Betrachten Sie nun den folgenden DFA.*

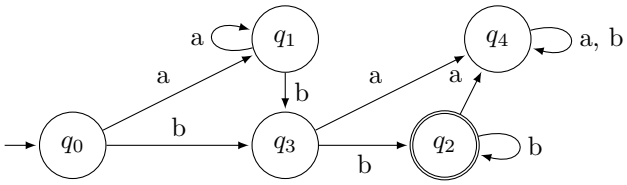


Minimieren Sie diesen DFA, indem Sie

1. *einen Trap State ergänzen, um den Automaten vollständig zu machen;*
2. *für den resultierenden Automaten die entsprechende 4×4 Matrix angeben und angeben, welche Klassen von äquivalenten Zuständen sich ergeben.*
3. *Geben Sie den resultierenden minimierten Automaten an.*

Lösung

1.



1 Pkt

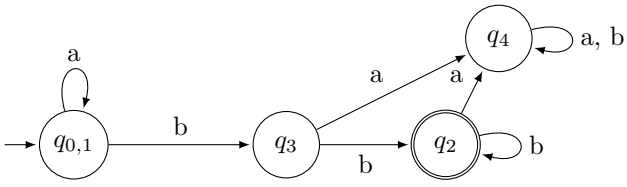
2.

	0	1	2	3
4	x	x	x	x
3	x	x	x	
2	x	x		
1				

Äquivalent sind q_0 und q_1.

5 Pkte

3.



2 Pkte

Aufgabe 4 *(5+3 Pkte)*

1. *Welche Sprache wird jeweils von den folgenden regulären Ausdrücken denotiert?*
   (a) $ab|\varepsilon c\varepsilon$ *(b)* $(ab|bb)^*$ *(c)* $((\emptyset abc^*)|\varepsilon b|(\varepsilon|a)b)^+$
2. *Geben Sie für jede der folgenden Sprachen einen regulären Ausdruck an, der diese denotiert.*

(a) $\{a, bc, \varepsilon\}$

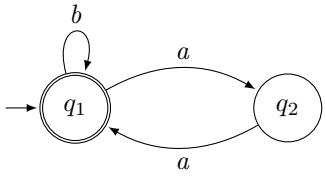
(b) $\{w \in \{a,b\}^* \mid$ *w hat eine Länge* $3 \cdot n$ *für ein* $n \geq 2\}$

Lösung:

1. (a) $L(ab|\varepsilon c\varepsilon) = \{ab, c\}$ 1 Pkt
   (b) $L((ab|bb)^*) = \{w \in \{a,b\}^* |\ w = 2n$ für ein $n \geq 0$ und jedes a in w ist notwendig von einem b gefolgt$\}$ 2 Pkte
   (c) $L(((\emptyset abc^*)|\varepsilon b|(\varepsilon|a)b)^+) = L((b|ab)^*) = \{w \in \{a,b\}^+ |$ jedes a in w ist notwendig von einem b gefolgt$\}$ 2 Pkte
2. (a) $(a|bc|\varepsilon)$ 1 Pkt
   (b) $(a|b)(a|b)(a|b)((a|b)(a|b)(a|b))^+$ 2 Pkte

Aufgabe 5 *(6 Pkte)*

Betrachten Sie nun den folgenden DFA:



Berechnen Sie den regulären Ausdruck, der die von diesem Automaten akzeptierte Sprache charakterisiert. Wenden Sie den Algorithmus aus der Vorlesung an, indem Sie rekursiv den Ausdruck $r^2_{1,1}$ *berechnen. Die Formel hierzu ist* $r^2_{1,1} = r^1_{1,1}|r^1_{1,2}(r^1_{2,2})^* r^1_{2,1}$.

Lösung:

$r^2_{1,1} = r^1_{1,1}|r^1_{1,2}(r^1_{2,2})^* r^1_{2,1}$

$r^1_{1,1} = r^0_{1,1}|r^0_{1,1}(r^0_{1,1})^* r^0_{1,1} = (b|\varepsilon)|(b|\varepsilon)^+(b|\varepsilon) = b^*$ 1 Pkt

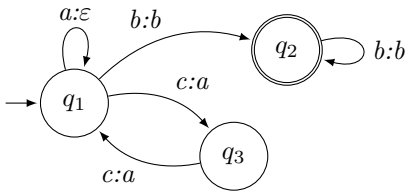
$r^1_{1,2} = r^0_{1,2}|r^0_{1,1}(r^0_{1,1})^* r^0_{1,2} = a|(b|\varepsilon)^+ a = b^*a$ 1 Pkt

$r^1_{2,2} = r^0_{2,2}|r^0_{2,1}(r^0_{1,1})^* r^0_{1,2} = \varepsilon|a(b|\varepsilon)^* a = \varepsilon|ab^*a$ 2 Pkte

$r^1_{2,1} = r^0_{2,1}|r^0_{2,1}(r^0_{1,1})^* r^0_{1,1} = a|a(b|\varepsilon)^+ = ab^*$ 1 Pkt

$r^2_{1,1} = b^*|b^*a(\varepsilon|ab^*a)^*ab^* = b^*|b^*a(ab^*a)^*ab^* = b^*|b^*(aab^*)^+ = b^*(aab^*)^*$ 1 Pkt

Aufgabe 6 *(4 Pkte) Betrachten Sie folgenden FST:*



1. *Worauf bilden der Automat die Eingaben accccaabb und aaaaabbbbb ab?*
2. *Welche Eingabewörter akzeptiert der Automat und wie transformiert er sie?*

Lösung:

1. Die Bilder sind *aaaabb* und *bbbbb*. 1 Pkt

2. Der Automat akzeptiert $L((a|cc)^*b^+)$, wobei alle as in der Eingabe gelöscht, alle cs durch a ersetzt und alle bs unverändert in die Ausgabe übernommen werden. 3 Pkte

Aufgabe 7 *(6 Pkte)*

Nehmen Sie an, wir wollen ein Bigram-Sprachmodell mit Laplace-Smoothing erstellen. Unsere Trainingsdaten sind nach Behandlung von unbekannten Wörtern (Zeichen `<UNK>`*) die folgenden zwei Sätze:*

`<s>` `<UNK>` *a a a* `<UNK>` *a* `</s>`
`<s>` *a a a a b a b a a b* `</s>`

1. *Wie sehen die folgenden Bigram-Wahrscheinlichkeiten in diesem Modell aus?*
   (a) $P(a|$`<s>`$)$ *(b)* $P($`<UNK>`$|a)$ *(c)* $P(a|$`<UNK>`$)$ *(d)* $P($`</s>`$|a)$
2. *Wie ist die Wahrscheinlichkeit und die Perplexity des folgenden Eingabesatzes?*

   `<s>` *a c a* `</s>`

   (Die Berechnung muss angegeben werden, der Wert muss jedoch nicht ausgerechnet werden.)

Lösung:

1. (a) $P(a|$`<s>`$) = \frac{1+1}{2+4} = \frac{1}{3}$ (b) $P($`<UNK>`$|a) = \frac{1+1}{11+4} = \frac{2}{15}$ (c) $P(a|$`<UNK>`$) = \frac{2+1}{2+4} = \frac{1}{2}$
   (d) $P($`</s>`$|a) = \frac{1+1}{11+4} = \frac{2}{15}$

   4 Pkte
2. Wahrscheinlichkeit $\frac{1}{3} \cdot \frac{2}{15} \cdot \frac{1}{2} \cdot \frac{1}{15} = \frac{2}{675} \approx 0.003$ 1 Pkt

   Perplexity $\sqrt[4]{\frac{675}{2}} \approx 4.286$ 1 Pkt

Aufgabe 8 *(8 Pkte)*

Nehmen Sie an, Sie haben einen HMM-POS Tagger, unter anderem mit Zuständen N und V und folgenden Wahrscheinlichkeiten:

Emissionswahrscheinlichkeiten:

$P(time|\,V) = 1 \cdot 10^{-3}$ $P(flies|\,V) = 2 \cdot 10^{-3}$
$P(time|N) = 4 \cdot 10^{-3}$ $P(flies|N) = 3 \cdot 10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für time *und* flies *seien* 0.

Übergangswahrscheinlichkeiten:

$P(N|start) = 2 \cdot 10^{-1}$ $P(N|N) = 1 \cdot 10^{-1}$ $P(N|\,V) = 2 \cdot 10^{-1}$
$P(\,V|start) = 1 \cdot 10^{-1}$ $P(\,V|N) = 4 \cdot 10^{-1}$ $P(\,V|\,V) = 1 \cdot 10^{-1}$
$P(end|N) = 1 \cdot 10^{-1}$ $P(end|\,V) = 2 \cdot 10^{-1}$

1. *Geben Sie die Viterbi Matrix an, die sich bei diesen Wahrscheinlichkeiten für die Eingabe* time flies *ergibt. Es reicht, die Einträge anzugeben, die* $\neq 0$ *sind. Geben Sie für jedes Feld Ihren Rechenweg an.*
2. *Was ist die beste POS-Tag Sequenz, die sich aus der Matrix für* time flies *ergibt?*

Lösung:

1.

end			$128 \cdot 10^{-9}$, V
N	$8 \cdot 10^{-4}$, start	$24 \cdot 10^{-8}$, N	
V	$1 \cdot 10^{-4}$, start	$64 \cdot 10^{-8}$, N	
	1 time	2 flies	

time, N: $P(\text{time}|\text{N}) \cdot P(\text{N}|\text{start}) = 4 \cdot 2 \cdot 10^{-4} = 8 \cdot 10^{-4}$ 1 Pkt

time, V: $P(\text{time}|\text{V}) \cdot P(\text{V}|\text{start}) = 1 \cdot 1 \cdot 10^{-4}$ 1 Pkt

flies, N: $\max\{8 \cdot 10^{-4} \cdot P(\text{N}|\text{N}) \cdot P(\text{flies}|\text{N})$ (Vorgänger N) ,
$1 \cdot 10^{-4} \cdot P(\text{N}|\text{V}) \cdot P(\text{flies}|\text{N})$ (Vorgänger V)$\}$
$= \max\{8 \cdot 1 \cdot 3 \cdot 10^{-8}, 1 \cdot 2 \cdot 3 \cdot 10^{-8}\} = 24 \cdot 10^{-8}$ 2 Pkte

flies, V: $\max\{8 \cdot 10^{-4} \cdot P(\text{V}|\text{N}) \cdot P(\text{flies}|\text{V})$ (Vorgänger N) ,
$1 \cdot 10^{-4} \cdot P(\text{V}|\text{V}) \cdot P(\text{flies}|\text{V})$ (Vorgänger V)$\}$
$= \max\{8 \cdot 4 \cdot 2 \cdot 10^{-8}, 1 \cdot 1 \cdot 2 \cdot 10^{-8}\} = 64 \cdot 10^{-8}$ 2 Pkte

q_F: $\max\{24 \cdot 10^{-8} \cdot 1 \cdot 10^{-1}$ Vorgänger N $, 64 \cdot 10^{-8} \cdot 2 \cdot 10^{-1}$ Vorgänger V$\}$ 1 Pkt

2. Die beste POS-Tag Folge ist demnach N V. 1 Pkt