Sommersemester 2018 29.05.2018

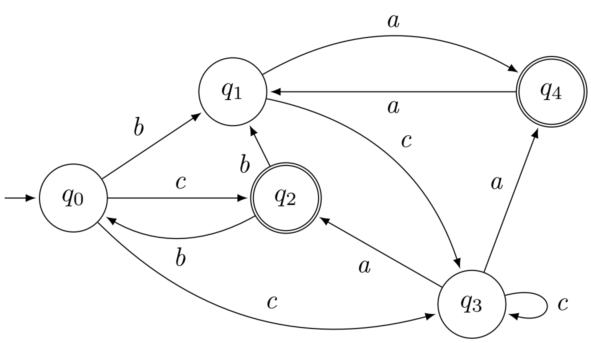
Einführung in die Computerlinguistik
Zwischenklausur

Laura Kallmeyer

Heinrich-Heine-Universität Düsseldorf

Erlaubte Hilfsmittel: Eine Din-A4 Seite mit Notizen.

Aufgabe 1 *(4 Pkte) Betrachten Sie den folgenden NFA:*

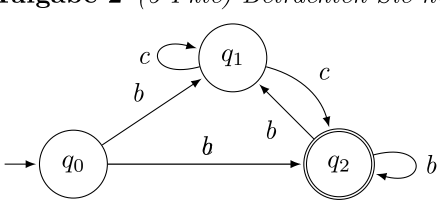


1. *Ist dieser Automat ein DFA, also deterministisch? Begründen Sie Ihre Antwort.*
2. *δ sei die Übergangsfunktion dieses Automaten. Berechnen Sie:*
   (i) $\delta(q_0, c)$ (ii) $\delta(q_3, b)$
3. *$\hat{\delta}$ sei die reflexive transitive Hülle von δ wie in der Vorlesung definiert. Berechnen Sie die folgenden Werte:*
   (i) $\hat{\delta}(q_0, \varepsilon)$ (ii) $\hat{\delta}(q_1, ac)$ (ii) $\hat{\delta}(q_1, aaca)$

Lösung:

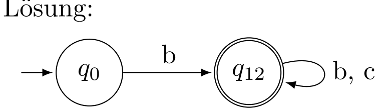
1. Nein, da z.B. q_3 zwei ausgehende Kanten mit gleichem Label (nämlich a) hat. 1 Pkt
2. (i) $\delta(q_0, c) = \{q_2, q_3\}$ (ii) $\delta(q_3, b) = \emptyset$ 1 Pkt
3. (i) $\hat{\delta}(q_0, \varepsilon) = \{q_0\}$ (ii) $\delta(q_1, ac) = \emptyset$ (ii) $\hat{\delta}(q_1, aaca) = \{q_2, q_4\}$ 2 Pkte

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

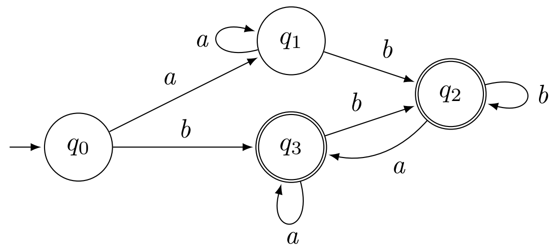


Konstruieren Sie einen äquivalenten DFA.

Lösung:



Aufgabe 3 *(6 Pkte) Betrachten Sie nun den folgenden DFA.*

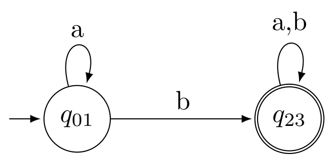


Minimieren Sie diesen DFA, indem Sie die entsprechende 3×3 Matrix angeben und anschließend äquivalente Zustände zusammenfassen. Geben Sie die Matrix und den resultierenden minimierten Automaten an.

Lösung

	0	1	2
3	x	x	
2	x	x	
1			

4 Pkte



2 Pkte

Aufgabe 4 *(5+5 Pkte)*

1. *Welche Sprache wird jeweils von den folgenden regulären Ausdrücken denotiert?*

   (a) $b|\varepsilon$ *(b)* $(c|ab|b)^+$ *(c)* $((\varepsilon|a)b|(\varepsilon|\emptyset)c)^+$

2. *Geben Sie für jede der folgenden Sprachen einen regulären Ausdruck an, der diese denotiert.*

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

   (b) $\{w \in \{a,b\}^* \mid$ *w hat eine Länge* $\geq 2\}$

   (c) $\{w \in \{a,b\}^* \mid$ *es dürfen keine zwei oder mehr b hintereinander stehen ohne as dazwischen*$\}$

Lösung:

1. (a) $L(b|\varepsilon) = \{b, \varepsilon\}$ 1 Pkt

   (b) $L((c|ab|b)^+) = \{w \in \{a,b,c\}^+ |$ jedes a in w ist notwendig von einem b gefolgt$\}$ 2 Pkte

   (c) $L(((\varepsilon|a)b|(\varepsilon|\emptyset)c)^+) = L((c|ab|b)^+)$ 2 Pkte

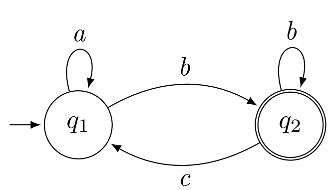
2. (a) $(a|\varepsilon)$ 1 Pkt

   (b) $(a|b)(a|b)^+$ 1 Pkt

   (c) $(a|ba)^*(\varepsilon|b)$ 3 Pkte

Aufgabe 5 *(8 Pkte)*

Betrachten Sie nun den folgenden DFA:



1. *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 $r^2_{1,2}$ berechnen. Die Formel hierzu ist $r^2_{1,2} = r^1_{1,2}|r^1_{1,2}(r^1_{2,2})^*r^1_{2,2}$.*

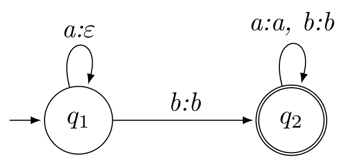
2. *Geben Sie eine rechtslineare Grammatik an, die genau die Sprache erzeugt, die von diesem Automaten akzeptiert wird.*

Lösung:

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

2. $G = \langle\{S, A\}, \{a, b, c\}, P, S\rangle$ mit folgenden Produktionen in P:
   $S \to aS \mid bA, A \to cS \mid bA \mid \varepsilon$ 3 Pkte

Aufgabe 6 *(3 Pkte) Betrachten Sie folgenden FST:*



1. *Worauf bilden der Automat die Eingaben bab und aabbab ab?*

2. *Welche Eingabewörter akzeptiert der Automat und wie transformiert er sie?*

Lösung:

1. Die Bilder sind *bab* und *bbab*. 1 Pkt

2. Der Automat akzeptiert beliebige Folgen von *a*s und *b*s, die mindestens ein *b* enthalten.

   Er löscht am Wortanfang alle *a*s, ab dem ersten Auftreten eines *b*s wird die Eingabe unverändert in die Ausgabe kopiert. 2 Pkte

Aufgabe 7 *(6 Pkte)*

Nehmen Sie an, wir wollen ein Bigram-Sprachmodell mit Laplace-Smoothing erstellen. Unsere Trainingsdaten sind die folgenden zwei Sätze:

<s> *a a b a b* </s>
<s> *a a b b b b a a b* </s>

1. *Wie sehen die folgenden Bigram-Wahrscheinlichkeiten in diesem Modell aus?*

   (a) $P(a|\texttt{<s>})$ *(b)* $P(b|a)$ *(b)* $P(a|b)$ *(b)* $P(\texttt{</s>}|a)$

2. *Wie ist die Wahrscheinlichkeit und die Perplexity des folgenden Eingabesatzes?*

   <s> *a b a* </s>

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

Lösung:

1. (a) $P(a|\texttt{<s>}) = \frac{3}{5}$ (b) $P(b|a) = \frac{5}{10} = \frac{1}{2}$ (b) $P(a|b) = \frac{3}{10}$ (b) $P(\texttt{</s>}|a) = \frac{1}{10}$

   4 Pkte

2. Wahrscheinlichkeit $\frac{3}{5} \cdot \frac{1}{2} \cdot \frac{3}{10} \cdot \frac{1}{10} = 9 \cdot 10^{-3} = 0.009$ 1 Pkt

   Perplexity $\sqrt[4]{\frac{1000}{9}}$ 1 Pkt

Aufgabe 8 *(8 Pkte)*

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

Emissionswahrscheinlichkeiten:

$P(the|Det) = 1$	$P(end|N) = 2 \cdot 10^{-3}$	$P(walks|N) = 1 \cdot 10^{-3}$
	$P(end|V) = 1 \cdot 10^{-3}$	$P(walks|V) = 3 \cdot 10^{-3}$

Alle anderen Emissionswahrscheinlichkeiten für end*,* the *und* walks *seien* 0*.*

Übergangswahrscheinlichkeiten:

$P(N|Det) = 5 \cdot 10^{-1}$	$P(N|N) = 1 \cdot 10^{-1}$	$P(N|V) = 2 \cdot 10^{-1}$
$P(V|Det) = 1 \cdot 10^{-1}$	$P(V|N) = 4 \cdot 10^{-1}$	$P(V|V) = 2 \cdot 10^{-1}$
	$P(Det|N) = 1 \cdot 10^{-1}$	$P(Det|V) = 2 \cdot 10^{-1}$

Angenommen, die Wahrscheinlichkeit, dass ein V am Satzanfang steht, ist $3 \cdot 10^{-1}$*, die, dass ein N am Satzanfang steht, ist* $2 \cdot 10^{-1}$ *und die, dass ein Det am Satzanfang steht, ist* $4 \cdot 10^{-1}$*.*

Die Wahrscheinlichkeit, dass auf ein N oder V ein Satzende folgt, ist jeweils $0.1 = 1 \cdot 10^{-1}$*.*

1. *Geben Sie die Viterbi Matrix an, die sich bei diesen Wahrscheinlichkeiten für die Eingabe* end the walks *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 als Ergebnis für* end the walks *ergibt?*

Lösung:

1.

q_F				$3 \cdot 10^{-9}$, N
N	$4 \cdot 10^{-4}$, q_0		$3 \cdot 10^{-8}$, Det	
V	$3 \cdot 10^{-4}$, q_0		$18 \cdot 10^{-9}$, Det	
Det		$6 \cdot 10^{-5}$, V		
	1 end	2 the	3 walks	

end, N: $P(\text{end}|\text{N}) \cdot 2 \cdot 10^{-1} = 2 \cdot 10^{-3} \cdot 2 \cdot 10^{-1}$ 1 Pkt

end, V: $P(\text{end}|\text{V}) \cdot 3 \cdot 10^{-1} = 1 \cdot 10^{-3} \cdot 3 \cdot 10^{-1}$ 1 Pkt

the, Det: $\max\{4 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 1 \text{ Vorgänger N }, 3 \cdot 10^{-4} \cdot 2 \cdot 10^{-1} \cdot 1 \text{ Vorgänger V}\}$ 2 Pkte

walks, N: $6 \cdot 10^{-5} \cdot 5 \cdot 10^{-1} \cdot 1 \cdot 10^{-3}$ 1 Pkt

walks, V: $6 \cdot 10^{-5} \cdot 1 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ 1 Pkt

q_F: $\max\{3 \cdot 10^{-8} \cdot 1 \cdot 10^{-1} \text{ Vorgänger N }, 18 \cdot 10^{-9} \cdot 1 \cdot 10^{-1} \text{ Vorgänger V}\}$ 1 Pkt

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