Residual Tree Languages

Home assignment to hand in before or on October 21, 2016.

October						
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Electronic versions (PDF only) can be sent by email to ⟨schmitz@lsv.ens-cachan.fr⟩, paper versions should be handed in on the 21st or put in my mailbox at LSV, ENS Cachan. **No delays**. The numbers in the margins next to exercises are indications of time and difficulty, not necessarily of the points you might earn answering them.

In this home assignment, we are interested in *residual languages* for tree languages over some finite ranked alphabet $\mathcal{F}$.

1 Bottom-up Residuals

As a reminder of the course (see pp. 35–36 of the lecture notes), recall that the *Myhill-Nerode congruence* of a tree language $L \subseteq T(\mathcal{F})$ is defined by

$$t \equiv_L t' \text{ iff } \forall C \in \mathcal{C}(\mathcal{F}) \,.\, C[t] \in L \Leftrightarrow C[t'] \in L \,.$$

By the Myhill-Nerode Theorem, a tree language L is recognisable if and only if $\equiv_L$ has finite index. The *minimal* bottom-up DFTA of L is then defined using the congruence classes of $\equiv_L$ as states.

Exercise 1 (Bottom-up Residuals)**.** Let $L \subseteq T(\mathcal{F})$ and $t \in T(\mathcal{F})$. The *bottom-up residual* of L by t is the set of all contexts C such that $C[t] \in L$:

$$t^{-1}L \stackrel{\text{def}}{=} \{C \in \mathcal{C}(\mathcal{F}) \mid C[t] \in L\} \,.$$

[1] 1. Show that L is recognisable if and only if $\{t^{-1}L \mid t \in T(\mathcal{F})\}$ is finite. Show that the minimal bottom-up DFTA for L has $|\{t^{-1}L \mid t \in T(\mathcal{F})\}|$ many states.

   Observe that $t \equiv_L t'$ if and only if $t^{-1}L = t'^{-1}L$ and apply the Myhill-Nerode Theorem.

2. Let $\mathcal{F} \stackrel{\text{def}}{=} \{f^{(2)}, a^{(0)}\}$ and consider for each $n > 0$ the language L_n of trees having at least one branch of length n:

$$L_n \stackrel{\text{def}}{=} \{t \in T(\mathcal{F}) \mid \exists p \in \mathrm{Pos}(t) \cap \mathbb{N}_{>0}^{n-1} \,.\, p1 \notin \mathrm{Pos}(t)\} \,.$$

[1] Give a NFTA for L_n with $n+1$ states.

In bottom-up rewriting notation, where $1 \leq i < n$:

$$a \to q_n \qquad f(q_*, q_{i+1}) \to q_i \qquad f(q_{i+1}, q_*) \to q_i$$
$$a \to q_* \qquad f(q_*, q_*) \to q_*$$

with accepting state q_1. The automaton guesses the leaf of a branch of length n, and recognises the branch using successive states $q_n, \ldots, q_1$; q_* acts as a universal state and recognises the remainder of the tree.

[2] 3. Show that the minimal bottom-up DFTA for L_n has at least 2^{n-1} states. You can consider for this the family of trees $t_K \stackrel{\text{def}}{=} t_{K,2}$ for $K \subseteq \{2, \ldots, n\}$ defined by

$$t_{K,n+1} \stackrel{\text{def}}{=} f(a,a) \qquad t_{K,i} \stackrel{\text{def}}{=} \begin{cases} f(t_{K,i+1}, a) & \text{if } i \in K \\ f(t_{K,i+1}, t_{K,i+1}) & \text{otherwise.} \end{cases}$$

The trees t_K have a branch of length $i \in \{2, \ldots, n\}$ if and only if $i \in K$. Consider the family of contexts C_i for $0 \leq i \leq n-2$ defined by

$$C_0 \stackrel{\text{def}}{=} \square \qquad C_{i+1} \stackrel{\text{def}}{=} f(C_i, a)\ .$$

Then all the branches in C_i are of length at most i and its hole is at depth exactly i. Therefore, $C_i[t_K]$ has a branch of length n if and only if $n-i \in K$.

Now, we wish to show that $t_K^{-1} L_n \neq t_{K'}^{-1} L_n$ for all $K \neq K'$, which will show that L_n has at least 2^{n-1} residuals. Indeed, wlog. $K \not\subseteq K'$, hence there exists $i \in K \setminus K'$, which means that $C_{n-i} \in t_K^{-1} L_n$ but $C_{n-i} \notin t_{K'}^{-1} L_n$.

2 Top-down Residuals

Exercise 2 (Top-down Residuals)**.** Let $L \subseteq T(\mathcal{F})$ and $C \in \mathcal{C}(\mathcal{F})$. The *top-down residual* of L by C is the set of all trees t such that $C[t] \in L$:

$$C^{-1}L \stackrel{\text{def}}{=} \{t \in T(\mathcal{F}) \mid C[t] \in L\}\ .$$

[3] 1. Show that if L is recognisable, then $|\{C^{-1}L \mid C \in \mathcal{C}(\mathcal{F})\}|$ is finite.

Let us consider the equivalence relation between contexts $C \approx_L C'$ if and only if $C^{-1}L = C'^{-1}L$, which is if and only if $\forall t \in T(\mathcal{F})\ .\ C[t] \in L \Leftrightarrow C'[t] \in L$.

If L is recognisable, then there is a complete bottom-up DFTA $\mathcal{A}$ with $L = L(\mathcal{A})$; let $n \stackrel{\text{def}}{=} |Q|$ be its number of states. Define another equivalence relation between contexts by $C \approx_{\mathcal{A}} C'$ if and only if $\forall q, q' \in Q\ .\ C[q] \to^* q' \Leftrightarrow C'[q] \to^* q'$. This equivalence relation has index at most n^n the number of functions from Q to Q (q' is uniquely determined).

Finally, observe that $C \approx_{\mathcal{A}} C'$ implies $C \approx_L C'$: for any $t \in T(\mathcal{F})$, there is a unique $q \in Q$ with $t \to^* q$, and then a unique $q' \in Q$ such that $C[t] \to^* C[q] \to^* q'$ and $C'[t] \to^* C'[q] \to^* q'$. Thus $C[t] \in L$ if and only if $C'[t] \in L$ if and only if $q' \in Q_f$. Hence $\approx_L$ has at most n^n equivalence classes.

[3] 2. Show that $|\{C^{-1}L_n \mid C \in \mathcal{C}(\mathcal{F})\}| = n+2$ for L_n defined in Exercise 1.

Let us first define the set of contexts with a branch of length n, not ending in the hole:

$$\mathcal{C}_* \stackrel{\text{def}}{=} \{C \in \mathcal{C}(\mathcal{F}) \mid \exists p \in \text{Pos}(C) \cap \mathbb{N}_{>0}^{n-1} \,.\, p1 \notin \text{Pos}(t) \wedge C(p) \neq \Box\} \,.$$

Then for all $t \in T(\mathcal{F})$ and $C \in \mathcal{C}_*$, $C[t] \in L_n$, hence all those contexts are equivalent for $\approx_{L_n}$.

Note that, if $C \notin \mathcal{C}_*$, the only possibility for $C[t]$ to be in L_n is then for the hole of C to be at depth at most n: the set of contexts that do not satisfy this is

$$\mathcal{C}_\infty \stackrel{\text{def}}{=} \{C \notin \mathcal{C}_* \mid \exists p \in Pos(C) \cap \mathbb{N}_{>0}^{\geq n} \,.\, C(p) = \Box\}$$

and for all $t \in T(\mathcal{F})$ and $C \in \mathcal{C}_\infty$, $C[t] \notin L_n$, hence all those contexts are equivalent for $\approx_{L_n}$ but not equivalent to any C from $\mathcal{C}_*$.

We further partition the remaining contexts depending on the depth of their hole: for all $0 \leq i < n$

$$\mathcal{C}_i \stackrel{\text{def}}{=} \{C \notin \mathcal{C}_* \mid \exists p \in Pos(C) \cap \mathbb{N}_{>0}^{i} \,.\, C(p) = \Box\}$$

is the set of contexts C such that $C[t] \in L_n$ if and only if $t \in L_{n-i}$. For each i, all those contexts are equivalent for $\approx_{L_n}$. Furthermore, consider $C_i \in \mathcal{C}_i$ and

- $C \in \mathcal{C}_*$: then any tree t with all its branches of length $> n$ has $C_i[t] \notin L_n$ but $C[t] \in L_n$, hence $C_i \not\approx_{L_n} C$;
- $C \in \mathcal{C}_\infty$: then any tree $t \in L_{n-i}$ is such that $C_i[t] \in L$ but $C[t] \notin L$, hence $C_i \not\approx_{L_n} C$;
- $C_j \in \mathcal{C}_j$ with $i \neq j$: then any tree $t \in L_{n-i} \setminus L_{n-j}$ (the latter set is not empty) is such that $C_i[t] \in L_n$ but $C_j[t] \notin L_n$, hence $C_i \not\approx_{L_n} C_j$.

Hence $\mathcal{C}_*, \mathcal{C}_\infty, \mathcal{C}_0, \ldots, \mathcal{C}_{n-1}$ is the partition defined by $\approx_{L_n}$ over $\mathcal{C}(\mathcal{F})$, with $n+2$ classes.

Notation. From this point on, we shall use top-down rewriting notations: transitions $(q, f^{(k)}, q_1, \ldots, q_k)$ are seen as rewriting rules $q \to f(q_1, \ldots, q_k)$. We also write '$I$' for the set of accepting states of a NFTA.

Exercise 3 (Context Determinististic Automata)**.** Let $\mathcal{A}$ be a NFTA $\mathcal{A} = \langle Q, \mathcal{F}, I, \delta\rangle$ and $C \in \mathcal{C}(\mathcal{F})$. We define

$$Q_C \stackrel{\text{def}}{=} \{q \in Q \mid \exists q_f \in I \,.\, q_f \to^* C[q]\}\ ,$$

as the set of states visited by C from some accepting state. We say that a NFTA is *context deterministic* if for all $C \in \mathcal{C}(\mathcal{F})$, $|Q_C| \leq 1$. (Note that this implies in particular $|I| \leq 1$ when considering $C = \square$ in the definition.)

[1] 1. Show that, if $\mathcal{A}$ is top-down deterministic, then it is context deterministic.

We proceed by induction on the depth of $\square$ in C. For the base case, if $C = \square$, then $Q_C = I$ and has cardinal at most one by definition of a top-down deterministic automaton. For the induction step, let $C = C'[f^k(t_1, \ldots, t_k)]$ with $t_i = \square$ for some $1 \leq i \leq k$. By induction hypothesis, $|Q_{C'}| \leq 1$, which means that there is at most one state q such that $q_f \to^* C'[q]$. Assuming q exists, since $\mathcal{A}$ is deterministic, there is at most one transition $q \to f^{(k)}(q_1, \ldots, q_k)$, and assuming this transition exists, $Q_C = \{q_i\}$.

[1] 2. Give an example of a context deterministic NFTA, which recognises a language not accepted by any top-down deterministic tree automaton.

The usual example works: $L = \{f(a, b), f(b, a)\}$ is recognised by the NFTA

$$q_f \to f(q_a, q_b) \qquad q_a \to a \qquad q_f \to f(q_b, q_a) \qquad q_b \to b$$

with accepting state q_f. The only five contexts with $Q_C \neq \emptyset$ are $Q_\square = \{q_f\}$, $Q_{f(\square,a)} = Q_{f(a,\square)} = \{q_b\}$ and $Q_{f(\square,b)} = Q_{f(b,\square)} = \{q_a\}$.

3. Assume we work with binary trees, i.e. that $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_2$. A tree language $L \subseteq T(\mathcal{F})$ is *homogeneous* if for all $C \in \mathcal{C}(F)$, $f \in \mathcal{F}_2$, and $t_1, t_2, t'_1, t'_2 \in T(\mathcal{F})$, whenever $C[f(t_1, t_2)] \in L$, $C[f(t_1, t'_2)] \in L$, and $C[f(t'_1, t_2)] \in L$, then $C[f(t'_1, t'_2)] \in L$.

[1] Show that, if L is recognised by a context deterministic NFTA, then L is homogeneous.

Assume $C[f(t_1, t_2)] \in L$, $C[f(t_1, t'_2)] \in L$, and $C[f(t'_1, t_2)] \in L$. This means that $Q_C = \{q\}$ for some state q. This gives rise to four interesting contexts, where $q_1 \to^* t_1$, $q_2 \to^* t_2$, $q'_1 \to^* t'_1$, and $q'_2 \to^* t'_2$:

$$Q_{C[f(\square,t_2)]} = \{q_1, q'_1\} \quad Q_{C[f(t_1,\square)]} = \{q_2, q'_2\} \quad Q_{C[f(\square,t'_2)]} = \{q_1\} \quad Q_{C[f(t'_1,\square)]} = \{q_2\}\ .$$

Since $\mathcal{A}$ is context deterministic, $q_1 = q'_1$ and $q_2 = q'_2$. In particular, we have the reduction $q_f \to^* C[q] \to^* C[f(t'_1, q_2)] \to^* C[f(t'_1, t'_2)]$, the latter being therefore in L.

[1] 4. For which n is L_n from Exercise 1 context deterministic?

It is context deterministic if and only if $n \leq 1$. Indeed, $L_1 = \{a\}$ is context deterministic, since it is recognised by a top-down deterministic automaton with unique transition $q_a \to a$.

Furthermore, by the previous question, it suffices to show that L_n for $n > 1$ is not homogeneous. Consider $t \in L_{n-1}$ and t' with all its branches of length at least n. Then $f(t,t)$, $f(t,t')$, and $f(t',t)$ are all in L_n. However, $f(t',t')$ is not.

Exercise 4 (Top-down Residual Automata)**.** Let $L = L(\mathcal{A})$ for a NFTA $\mathcal{A} = \langle Q, \mathcal{F}, I, \delta\rangle$ and $q \in Q$. Define the *state language* of q as

$$L_q \stackrel{\text{def}}{=} \{t \in T(\mathcal{F}) \mid q \to^* t\} .$$

We say that a NFTA is *top-down residual* if, for all $q \in Q$, there exists $C \in \mathcal{C}(\mathcal{F})$ such that $L_q = C^{-1}L$.

[1] 1. Show that top-down residuals of recognisable languages are unions of state languages, more exactly that for all $C \in \mathcal{C}(\mathcal{F})$,

$$C^{-1}L = \bigcup_{q \in Q_c} L_q .$$

For the direct inclusion, let t be such that $C[t] \in L$. Then there is a reduction $q_f \to^* C[q] \to^* C[t]$ for some $q \in Q$ and $q_f \in I$. Thus $t \in L_q$ and $q \in Q_C$.

Conversely, let $t \in L_q$ for some $q \in Q_C$. Thus $q_f \to^* C[q] \to^* C[t]$, hence $t \in C^{-1}L$.

[1] 2. Assume $\mathcal{A}$ is accessible[1] and context deterministic. Show that $\mathcal{A}$ is top-down residual.

Since $\mathcal{A}$ is accessible, for all $q \in Q$, there exists $C \in \mathcal{C}(\mathcal{F})$ such that $q \in Q_C$, and since $\mathcal{A}$ is context deterministic, $Q_C = \{q\}$. Hence for all q, there exists C with $C^{-1}L = L_q$ by the previous question.

[1] 3. Show that there exists a top-down residual NFTA for L_n of Exercise 1 with $n+1$ states.

It suffices to show that the NFTA provided in Question 2 of Exercise 1 is top-down residual. Observe that $L_{q_*} = C^{-1}L_n$ for any $C \in \mathcal{C}_*$ as defined in Question 2 of Exercise 2. Then, for all $1 \leq i \leq n$, $L_{q_i} = L_{n-i} = C_{i-1}^{-1}L_n$ for any $C_{i-1} \in \mathcal{C}_{i-1}$ as defined in Question 2 of Exercise 2.

[1] 4. Show that the language $L = \{f(a,b), f(a,c), f(b,a), f(b,c), f(c,a), f(c,b)\}$ is not recognised by any top-down residual automaton.

[1] Recall that $\mathcal{A}$ is *trim* if it is co-accessible ($\forall q \in Q, \exists t \in T(\mathcal{F}) \,.\, q \to^* t$) and accessible ($\forall q \in Q, \exists q_f \in I, \exists C \in \mathcal{C}(\mathcal{F}) \,.\, q_f \to^* C[q]$).

The top-down residuals are $\Box^{-1}L = L$ and

$$f(a,\Box)^{-1}L = f(\Box,a)^{-1}L = \{b,c\}$$
$$f(b,\Box)^{-1}L = f(\Box,b)^{-1}L = \{a,c\}$$
$$f(c,\Box)^{-1}L = f(\Box,c)^{-1}L = \{a,b\}$$

Assume L is recognised by a top-down residual automaton $\mathcal{A}$. Consider a run for $f(a,b)$: we have three states with transitions $q_0 \to f(q_1,q_2)$, $q_1 \to a$, and $q_2 \to b$. Since $\mathcal{A}$ is top-down residual, as L_{q_1} contains a it can be either $\{a,b\}$ or $\{a,c\}$. Similarly, L_{q_2} can be either $\{b,c\}$ or $\{a,b\}$. Hence $\mathcal{A}$ would also recognise one of $f(a,a)$, $f(b,b)$, or $f(c,c)$.

Exercise 5 (Minimal Top-down Residual Automata)**.** Let L be a language recognised by some top-down residual automaton $\mathcal{A} = \langle Q, \mathcal{F}, I, \delta\rangle$. We say that a top-down residual $C^{-1}L$ is *composite* if it is the union of the residuals it contains, i.e. if

$$C^{-1}L = \bigcup_{C'^{-1}L \subsetneq C^{-1}L} C'^{-1}L\ .$$

We call it *prime* if it is not composite.

[1] 0. Show that any residual $C^{-1}L$ is a finite union of prime residuals.

We prove this by induction on the strict ordering $\subsetneq$ over the set of residuals of L; as $\{C^{-1}L \mid C \in \mathcal{C}(\mathcal{F})\}$ is finite by Question 1 of Exercise 2, this is a well-founded ordering. For the base case, if $C^{-1}L = \emptyset$, then it is an empty union of prime residuals. For the induction step, either $C^{-1}L$ is prime, or it is composite and by induction hypothesis every $C'^{-1}L \subsetneq C^{-1}L$ is a finite union of prime residuals; hence in both cases $C^{-1}L$ is a finite union of prime residuals.

[1] 1. Show that, if $C^{-1}L$ is prime, then there exists $q \in Q$ such that $L_q = C^{-1}L$.

As seen in Question 1 of the previous exercise, $C^{-1}L = \bigcup_{q \in Q_C} L_q$. Since $\mathcal{A}$ is top-down residual, for each $q \in Q_C$, there exists C_q such that $L_q = C_q^{-1}L$. Hence $C^{-1}L = \bigcup_{q\in Q_C} C_q^{-1}L$. If $C_q^{-1}L \subsetneq C^{-1}L$ for all $q \in Q_C$, $C^{-1}L$ would be composite, hence there exists $q \in Q_C$ such that $C^{-1}L = L_q = C_q^{-1}L$.

2. We define now the *canonical* top-down residual automaton of L, which uses prime residuals as states: $\mathcal{A}_L \stackrel{\text{def}}{=} \langle Q_L, \mathcal{F}, I_L, \delta_L\rangle$ where

$$\begin{aligned}
Q_L &\stackrel{\text{def}}{=} \{C^{-1}L \in \mathcal{C}(\mathcal{F}) \mid C^{-1}L \text{ prime}\}\\
I_L &\stackrel{\text{def}}{=} \{C^{-1}L \in Q \mid C^{-1}L \subseteq L\}\\
\delta_L &\stackrel{\text{def}}{=} \{C^{-1}L \to a^{(0)} \mid C[a] \in L\}\\
&\quad \cup\ \{C^{-1}L \to f^{(k)}(C_1^{-1}L,\ldots,C_k^{-1}L) \mid \forall (t_1,\ldots,t_k) \in C_1^{-1}L \times \cdots \times C_k^{-1}L\,.\, C[f(t_1,\ldots,t_k)] \in L\}\ .
\end{aligned}$$

[4] Show that for all $C^{-1}L \in Q_L$, $C^{-1}L = L_{C^{-1}L}$ in $\mathcal{A}_L$.

Hint: Use the existence of $\mathcal{A} = \langle Q, \mathcal{F}, I, \delta\rangle$ top-down residual with $L = L(\mathcal{A})$.

Let us prove that for all $C^{-1}L \in Q_L$ and by induction on $t \in T(\mathcal{F})$, $t \in C^{-1}L$ if and only if $t \in L_{C^{-1}L}$, which will show that $C^{-1}L = L_{C^{-1}L}$.

For the base case $t = a^{(0)}$, assume first that $a \in C^{-1}L$, i.e. that $C[a] \in L$. Then $C^{-1}L \to a$ shows that $a \in L_{C^{-1}L}$. Conversely, if $C^{-1}L \to^* a$, then $C[a] \in L$ and $a \in C^{-1}L$.

For the induction step, consider $t = f^{(k)}(t_1, \ldots, t_k)$.

(a) Assume first that $t \in C^{-1}L$. We cannot use the induction hypothesis directly, because the residuals of $C[f^{(k)}(t_1, \ldots, t_{j-1}, \Box, t_{j+1}, \ldots, t_k)]$ might not be prime. We use instead the top-down residual automaton $\mathcal{A} = \langle Q, \mathcal{F}, I, \delta\rangle$. Since $C^{-1}L$ is prime, by the first question there exists $q \in Q$ such that $L_q = C^{-1}L$, hence $t \in L_q$. Thus there exists a transition $q \to f^{(k)}(q_1, \ldots, q_k)$ in δ with $t_j \in L_{q_j}$ for all $1 \le j \le k$. More generally, any tuple $(t'_1, \ldots, t'_k) \in L_{q_1} \times \cdots \times L_{q_k}$ is such that $f(t'_1, \ldots, t'_k) \in L_q = C^{-1}L$. For each $q_j \in Q$ of a top-down residual automaton, L_{q_j} is a residual:

- If L_{q_j} is a prime residual, there exists $C_j^{-1}L \in Q_L$ such that $t_j \in C_j^{-1}L$.
- Otherwise, L_{q_j} is composite; nevertheless, by Question 0 we can decompose it into prime residuals so that $t_j \in C_j^{-1}L \subsetneq L_{q_j}$ for some $C_j^{-1}L \in Q_L$.

Hence, for all $(t'_1, \ldots, t'_k) \in C_1^{-1}L \times \cdots \times C_k^{-1}L \subseteq L_{q_1} \times \cdots \times L_{q_k}$, $f(t'_1, \ldots, t'_k) \in L_q = C^{-1}L$, i.e. $C[f(t'_1, \ldots, t'_k) \in L]$. Therefore, $C^{-1}L \to f^{(k)}(C_1^{-1}L, \ldots, C_k^{-1}L)$ is a transition in δ_L. By induction hypothesis, $t_j \in C_j^{-1}L = L_{C_j^{-1}L}$ for all $1 \le j \le k$, and therefore $t \in L_{C^{-1}L}$.

(b) Assume conversely that $t \in L_{C^{-1}L}$. Hence $C^{-1}L \to f^{(k)}(C_1^{-1}L, \ldots, C_k^{-1}L) \in \delta_L$ with $C_j^{-1}L \to^* t_j$ for all $1 \le j \le k$. By induction hypothesis, each $t_j \in C_j^{-1}L$. By definition of δ_L, this implies $C[f(t_1, \ldots, t_k)] \in L$, hence $t \in C^{-1}L$ as desired.

[2] 3. Conclude that $L = L(\mathcal{A}_L)$, $\mathcal{A}_L$ is top-down residual, and has a minimal number of states among all possible top-down residual NFTA for L.

Consider the context $\Box$: $L = \Box^{-1}L$ is a union of prime residuals. Hence any $t \in L$ belongs to some prime $C^{-1}L \in Q_L$ with $C^{-1}L \subseteq L$, hence $C^{-1}L$ is in I_L. Since $C^{-1}L = L_{C^{-1}L}$ by the previous question, $t \in L(\mathcal{A}_L)$. Conversely, if $t \in L(\mathcal{A}_L)$, then there exists $C^{-1}L \in I_L$ such that $t \in L_{C^{-1}L} = C^{-1}L \subseteq L$.

By the previous question, $C^{-1}L = L_{C^{-1}L}$ implies that $\mathcal{A}_L$ is top-down residual.

Finally, by the first question, any top-down residual NFTA for L has at least one state per prime residual. Therefore, $\mathcal{A}_L$ is minimal.