Tree Automata and Applications

Exercise session 3 Solutions

Luc Lapointe
luc.lapointe@ens-paris-saclay.fr
home.lmf.cnrs.fr/LucLapointe/

Exercise 1 - Recognizing an abstract language

1. Let $\mathcal{E}$ be a finite set of linear terms on $T(\mathcal{F}, \mathcal{X})$. Prove that

$$\mathrm{Red}(\mathcal{E}) \stackrel{\text{def}}{=} \{C[t\sigma] \mid C \in \mathcal{C}(\mathcal{F}), t \in \mathcal{E}, \sigma \text{ ground substitution}\}$$

is recognizable.

Correct automaton

Let t be a linear term on $T(\mathcal{F}, \mathcal{X})$. The following NFTA recognizes $\mathrm{Red}(\{t\})$: $Q = \{q_\perp\} \cup \mathrm{Pos}(t)$, $F = \{\varepsilon\}$ and transitions as follow:

$$\begin{aligned}
f(q_1, \dots, q_n) &\longrightarrow q_\perp \text{ for all } f \in \mathcal{F}, q_1, \dots, q_n \in Q \\
q_\perp &\longrightarrow p \text{ for all } p \in \mathrm{Pos}(t) \text{ such that } t(p) \text{ is a variable} \\
f(p.1, \dots, p.n) &\longrightarrow p \text{ if } t(p) = f \\
f(q_1, \dots, q_n) &\longrightarrow \varepsilon \text{ for all } f \in \mathcal{F} \text{ and } q_1, \dots, q_n \in Q \\
&\qquad \text{such that there exists } i \in \{1, \dots, n\} \text{ such that } q_i = \varepsilon
\end{aligned}$$

2. Prove that if $\mathcal{E}$ contains only ground terms, then one can construct a DFTA recognizing $\mathrm{Red}(\mathcal{E})$ whose number of states is at most $n + 2$, where n is the number of nodes of $\mathcal{E}$.

Solution

Let $\mathrm{St}(\mathcal{E})$ be the set of all subterms of a term in $\mathcal{E}$. Then the following DFTA works: $Q = \{q_t \mid t \in \mathrm{St}(\mathcal{E})\} \cup \{q_\top, q_\perp\}$, $F = \{q_\top\}$ and transitions is the **union**, for all $f \in \mathcal{F}$, of:

$$f(q_1, \dots, q_n) \longrightarrow \begin{cases} q_{f(t_1, \dots, t_n)} \text{ if } \forall i, \exists t_i,\, q_i = t_i \text{ and } f(t_1, \dots, t_n) \in \mathrm{St}(\mathcal{E}) \setminus \mathcal{E} \\ q_\top \text{ if } \forall i, \exists t_i,\, q_i = t_i \text{ and } f(t_1, \dots, t_n) \in \mathcal{E} \\ q_\top \text{ if there is at least one } i \text{ such that } q_i = q_\top \\ q_\perp \text{ otherwise} \end{cases}$$

Scheme of the proof:

1. Prove by induction on the size of the terms that for all $t \in \mathrm{St}(\mathcal{E}) \setminus \mathcal{E}$, the language of q_t is $\{t\}$.

2. Prove by induction on the size n of the terms that

$$\begin{aligned} &\mathrm{Red}^{<n}(\mathcal{E}) \subset L(q_\top) \\ &\wedge T^{<n}(\mathcal{F}, \mathcal{X}) \setminus (\mathrm{Red}(\mathcal{E}) \cup \mathrm{St}(\mathcal{E})) \subset L(q_\bot) \end{aligned}$$

where $L^{<n}$ denotes the trees of size at most n.

3. Conclude that $L(q_\top) = \mathrm{Red}(\mathcal{E})$ because the automaton is deterministic.

Exercise 2 - Decision problems

Let $\mathcal{F}$ a signature. We consider the Ground Instance Intersection **(GII)** problem:
Input : t a term in $T(\mathcal{F}, \mathcal{X})$ and $\mathcal{A}$ a NFTA.
Question : Is there at least one ground instance of t accepted by $\mathcal{A}$?

1. Suppose that t is linear. Prove that **(GII)** is P-complete.
Hint: The emptiness problem for tree automata is P-complete.

Solution idea

In P: Use a construction similar to Exercise 1.1. This construction $\mathcal{A}_1$ contains $O(|t|^k)$ transitions, where k is the maximum arity of a symbol in $\mathcal{F}$, which is a constant because $\mathcal{F}$ is not an input of the problem. Build intersection with $\mathcal{A}$ in time linear in $|\mathcal{A}| * |\mathcal{A}_1|$, and test the non-emptiness in polynomial time.

P-hard: Reduction from the emptiness problem for tree automata, which is P-hard. Testing the emptiness of $\mathcal{A}$ is equivalent to testing **(GII)** on $\mathcal{A}$ and a variable, so the polynomial reduction mapping $\mathcal{A}$ to $x, \mathcal{A}$ works.

2. Suppose that $\mathcal{A}$ is deterministic. Prove that **(GII)** is NP-complete.

Solution idea

In NP: Guess for each variable an accessible state of $\mathcal{A}$. Verify that you can complete this to an accepting run by running the automaton. This verification can be achieved in polynomial time because the automaton is deterministic.

NP-hard: Reduction from SAT. Let $\mathcal{F} = \{\mathrm{not}(1), \mathrm{or}(2), \mathrm{and}(2), \bot(0), \top(0)\}$ and $\mathcal{A}_{\mathrm{SAT}}$ the DFTA with $Q = \{q_\top, q_\bot\}, F = \{q_\top\}$ and transitions as follow:

$$\bot \longrightarrow q_\bot \qquad \top \longrightarrow q_\top \qquad \mathrm{not}(q_\alpha) \longrightarrow q_{\neg\alpha}$$

$$\text{or}(q_\alpha, q_\beta) \longrightarrow q_{\alpha\vee\beta} \qquad \text{and}(q_\alpha, q_\beta) \longrightarrow q_{\alpha\wedge\beta}$$

The language of $\mathcal{A}_{\text{SAT}}$ is the set of closed valid formulae.

Let φ a CNF formula, $\varphi = \bigwedge_{i=1}^{n} c_i$ where c_i are clauses. Define t_{c_i} by induction on the size of c_i as expected, with propositions denoted as variables, and

$$t_\varphi \stackrel{\text{def}}{=} \text{and}(t_{c_1}, \text{and}(\dots \text{and}(c_{n-1}, t_{c_n}))).$$

φ is satisfiable if and only if a closed instance of t_φ is recognized by $\mathcal{A}_{\text{SAT}}$.

3. Prove that **(GII)** is EXPTIME-complete.
Hint: For the hardness, reduce the intersection non-emptiness problem (admitted to be EXPTIME-complete).

Solution idea

In EXP: For each labelling of t by states (exponentially many):
- Check that the labelling of every occurrence of a variable is an accessible state (in polynomial time)
- Check that the labelling corresponds to an accepting run (in polynomial time)
- For every variable x, let $\{q_1, \dots, q_n\}$ be the set of labellings of all occurrences of x. For $q \in Q$, let $\mathcal{A}_q$ be the NFTA obtained from $\mathcal{A}$ by changing the set of final states to $\{q\}$. Check that

$$L(\mathcal{A}_{q_1}) \cap \dots \cap L(\mathcal{A}_{q_n})$$

  is non-empty.

EXP-hard: Reduce intersection non-emptiness. Let

$$\mathcal{A}_k = (Q_k, \mathcal{F}, I_k, \Delta_k)_{k\in\{1,\dots,n\}}$$

a finite sequence of top-down NFTA. Note that we can transform a bottom-up NFTA to a top-down one in polynomial time. Assume that all the Q_k are disjoint. Define:
- $\mathcal{F}' = \mathcal{F} \cup \{h(n)\}$
- $t = h(x, \dots, x)$
- $\mathcal{A}' = (\bigsqcup Q_k \sqcup \{q_0\}, \mathcal{F}', \{q_0\}, \Delta' \sqcup \bigsqcup \Delta_k)$ where

$$\Delta' = \{q_0(h(x_1, \dots, x_n)) \longrightarrow h(q_1(x_1), \dots, q_n(x_n)) \mid q_k \in I_k\}$$

  then $L(\mathcal{A}_1) \cap \dots \cap L(\mathcal{A}_n) \neq \emptyset$ iff t has a closed instance in $L(\mathcal{A}')$.

4. Deduce that the following problem is decidable:
   Input : t a term in $T(\mathcal{F};\mathcal{X})$ and linear terms $t_1, \dots, t_n$.
   Question : Is there a ground instance of t which is not an instance of any t_i?

Solution

Use Exercise 1.1 to build an automaton that recognizes the set of ground instances of each of the t_i. Make it deterministic (e.g. through powerset construction), then complement it. Use Question 3 on t and this construction.

Exercise 3 - Residuals

Exercise from January 2023 exam.

For $\mathcal{F} = \{f(2), a(0)\}$ and $n > 0$, let L_n be the language of trees that have at least one branch of length exactly n, i.e.

$$L_n \stackrel{\text{def}}{=} \{t \in T(\mathcal{F}) \mid \exists p \in \text{Pos}(t) : |p| = n - 1 \land t(p) = a\}$$

e.g. $f(a, f(a, f(a, a))) \in L_3$ because it contains one branch of length 3 (as well as one of length 2 and two of length 4).

1. Give a bottom-up NFTA for L_n with $n + 1$ states.

Correct automaton

Each leaf guesses whether it is on a path of length n, and then starts counting from q_1, or on some other path. Formally, consider $Q = \{q, q_1, \dots, q_n\}$, $F = \{q_n\}$ and Δ containing, for all $i \in [\![1, \dots, n - 1]\!]$:

$$a \to q \qquad a \to q_1 \qquad f(q, q) \to q \qquad f(q_i, q) \to q_{i+1} \qquad f(q, q_i) \to q_{i+1}$$

2. Show that the minimal DFTA for L_n has at least 2^{n-1} states.
   Hint: For $I \subseteq \{2, \dots, n\}$, build a tree t_I that has a branch of length i iff $i \in I$.

Solution

The result is true for $n = 1$ as an automaton with no state can not accept anything.

Let $K \stackrel{\text{def}}{=} \{2, ..., n\}$. Define

$$t_I^{(i)} = \begin{cases} f\left(a, t_I^{(i+1)}\right) & \text{if } i \in I \\ f\left(t_I^{i+1}, t_I^{i+1}\right) & \text{if } i \in K \setminus I \\ f(a, a) & \text{otherwise} \end{cases}$$

then $t_I \stackrel{\text{def}}{=} t_I^2$ has a branch of length i iff $i \in I$.

Suppose that we have a DFTA $\mathcal{A}$ accepting L_n with fewer than 2^{n-1} states. Then there must exist two different sets $I, J \subseteq K$ and a state q of $\mathcal{A}$ such that $t_I \to_{\mathcal{A}}^* q$ and $t_J \to_{\mathcal{A}}^* q$. Let i be the maximal index in the symmetric difference of I and J, and w.l.o.g. suppose that $i \in I \setminus J$.

We now consider the family of contexts $C_0 = x_1$ and $C_{k+1} = f(t_\emptyset, C_k)$ for $k \geq 0$. Then $C_{n-i}[t_I] \in L_n$ but $C_{n-i}[t_J] \notin L_n$. However, they are either both accepted or both rejected by $\mathcal{A}$: contradiction. ■

Let $L \subseteq T(\mathcal{F})$ be a language of trees and $C \in \mathcal{C}(\mathcal{F})$ a context. The *residual* of L by C is defined as

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

We define $R(L) \stackrel{\text{def}}{=} \{C^{-1}L \mid C \in \mathcal{C}(\mathcal{F})\}$ as the set of residuals of L.

3. Show that if L is recognizable, then $|R(L)|$ is finite.

Solution

If L is recognizable, let $\mathcal{A} = \langle Q, \mathcal{F}, G, \Delta \rangle$ be a DFTA with n states accepting L, and let $L_q = \left\{t \in T(\mathcal{F}) \mid t \xrightarrow{*}_{\mathcal{A}}^{*} q\right\}$. For $C \in \mathcal{C}(\mathcal{F})$, let $Q_C = \left\{q \in Q \mid \exists q' \in G : C[q] \xrightarrow{*}_{\mathcal{A}}^{*} q'\right\}$. Then $C^{-1}L = \bigcup_{q \in Q_C} L_q$. Since $C^{-1}L$ is entirely determined by Q_C, we have $|R(L)| \leq 2^{|Q|}$.

Homework - Direct images of a homomorphism

Let $\mathcal{F} = \{f(2), g(1), a(0)\}$ and $\mathcal{F}' = \{f'(2), g(1), a(0)\}$. Let us consider the tree homomorphism h determined by $h_{\mathcal{F}}$ defined by

$$h_{\mathcal{F}}(f) = f'(x_1, x_2)$$
$$h_{\mathcal{F}}(g) = f'(x_1, x_1)$$
$$h_{\mathcal{F}}(a) = a.$$

1. Is $h(T(\mathcal{F}))$ recognizable?
2. Let $L = \{g^i(a) \mid i \geq 0\}$, then L is a recognizable tree language. Is $h(L)$ recognizable?