Efficient universality of shuffle regular expressions

LC

May 1, 2019

Abstract

We show that universality and inclusion of shuffle regular expressions is decidable. This is defined by reduction to a new automaton model called shuffle finite automata, for which a direct procedure showing decidability of universality is provided. We also provide efficient algorithms based on subsumption techniques. This has applications in checking containment of Relax NG specifications where the horizontal languages are represented as shuffle regular expressions.

1 Introduction

The introduction of the shuffle and the shuffle closure operation leads to undecidability of the inclusion problem.

2 Shuffle finite automata

A *shuffle finite automaton* is a tuple $(\Sigma, P, I, F, \Delta)$ where Σ is a finite alphabet, P is a finite set of states, of which those in $I \subseteq P$ are initial and in $F \subseteq P$ are final, and $\Delta \subseteq P \times \Sigma \times \mathcal{M}_{\text{fin}}(P)$ is a (nondeterministic) transition relation. We call a finite multiset $M \in \mathcal{M}_{\text{fin}}(P)$ a *multistate*, and we say that it is *final* if contains a final state. For a state $p \in P$, an input symbol $a \in \Sigma$, and multistate $M \in \mathcal{M}_{\text{fin}}(P)$, we write $p \xrightarrow{a} M$ instead of $(p, a, M) \in \Delta$. The language $\mathcal{L}(p) \subseteq \Sigma^*$ of words *accepted* by state p is defined by induction on the length of words: $\varepsilon \in \mathcal{L}(p)$ if $p \in F$, and $aw \in \mathcal{L}(p)$ if there exists a transition $p \xrightarrow{a} M$ s.t. $w \in \mathcal{L}(M)$. Here, for a multistate $M = \{\{q_1, \ldots, q_k\}\}$, we define

$$\mathcal{L}(M) = \bigcup \{\mathcal{L}(q_1) \odot \cdots \odot \mathcal{L}(q_k) \mid \{\{q_1, \ldots, q_k\}\} \in \mathcal{M}_{\text{fin}}(M), k \geqslant 1\}.$$

The transition relation $p \xrightarrow{a} M$ can be simultaneously extended to a relation $M \xrightarrow{w} N$ for multistates M, N and an arbitrary word $w \in \Sigma^*$, as follows: $M \xrightarrow{\varepsilon} N$ for a nonempty $N \leqslant M$, and $M \xrightarrow{aw} N$ if there exists a state $q \in M$ s.t. $q \xrightarrow{a} O$ and $N = M - \{\{q\}\} + O'$ for a nonempty $O' \leqslant O$. We have $w \in \mathcal{L}(p)$ if, and only if, $\{\{p\}\} \xrightarrow{w} M$ for an final multistate M.

Consider the following *forward simulation* relation $\sqsubseteq_{\mathsf{fw}} \subseteq P \times P$ on states: For every two states $p, q \in P$, if $p \sqsubseteq_{\mathsf{fw}} q$, then,

1. if $p \in F$, then $q \in F$, and
2. $\forall (p \xrightarrow{a} M) \cdot \exists (q \xrightarrow{a} N) \cdot M \sqsubseteq^*_{\mathsf{fw}} N$.

Here, we define $M \sqsubseteq^*_{\mathsf{fw}} N$ if there exists an *injection* $f : M \to N$ s.t., for every $r \in M$, we have $r \sqsubseteq f(r)$. We also write $M \sqsubseteq^{*,f}_{\mathsf{fw}} N$ when we want to emphasise the concrete injection in use. Subsumption is defined in order to imply language inclusion.

Lemma 1. *If $p \sqsubseteq_{\mathsf{fw}} q$, then $\mathcal{L}(p) \subseteq \mathcal{L}(q)$. Consequently, if $M \sqsubseteq^*_{\mathsf{fw}} N$, then $\mathcal{L}(M) \subseteq \mathcal{L}(N)$.*

Proof. We proceed by induction on the length of words. If $\varepsilon \in \mathcal{L}(p)$, then $p \in F$, and thus $q \in F$, which implies $\varepsilon \in \mathcal{L}(q)$. For the inductive step, let $aw \in \mathcal{L}(p)$. There exists a transition $p \xrightarrow{a} M$ s.t. $w \in \mathcal{L}(M)$. Consequently, there are states $\{\{r_1, \ldots, r_k\}\} \in \mathcal{M}_{\mathrm{fin}}(M)$ and words $w_1, \ldots, w_k \in \Sigma^*$ s.t. $w = w_1 \odot \cdots \odot w_k$ and $w_1 \in \mathcal{L}(r_1), \ldots, w_k \in \mathcal{L}(r_k)$. By the definition of subsumption, there is a transition $q \xrightarrow{a} N$ with $M \sqsubseteq^{*,f}_{\mathsf{fw}} N$ s.t. for every $r \in M$ we have $r \sqsubseteq_{\mathsf{fw}} f(r)$. In particular, $r_1 \sqsubseteq_{\mathsf{fw}} f(r_1), \ldots, r_k \sqsubseteq_{\mathsf{fw}} f(r_k)$, and thus, by induction hypothesis, $w_1 \in \mathcal{L}(f(r_1)), \ldots, w_k \in \mathcal{L}(f(r_k))$. Consequently, $w \in \mathcal{L}(f(r_1)) \odot \cdots \odot \cdots \mathcal{L}(f(r_k))$, which implies $w \in \mathcal{L}(N)$. Finally, $aw \in \mathcal{L}(q)$, as required. □

Consider the following *backward simulation* relation $\sqsubseteq_{\mathsf{bw}} \subseteq P \times P$ on states:

Computing simulations.

3 Universality

We show an algorithm to decide the universality problem for shuffle automata. Fix a shuffle automaton $\mathcal{A} = (\Sigma, P, F, \Delta)$. We perform a subset construction leading to a deterministic infinite-state automaton $\mathcal{B}$ recognising the same language.

Each *configuration* of $\mathcal{B}$ is a set

$$C = \{M_1, \ldots, M_k\}$$

of macrostates $M_1, \ldots, M_k \in \mathcal{M}_{\mathrm{fin}}(P)$. Let $P' := 2^{\mathcal{M}_{\mathrm{fin}}(P)}$ be the set of configurations. If the initial states of $\mathcal{A}$ are $I = \{q_1, \ldots, q_k\}$, then there is a single *initial configuration* in $\mathcal{B}$, namely

$$I' = \{C_I\}, \text{ with } C_I = \{\{\{q_1\}\}, \ldots, \{\{q_k\}\}\},$$

and *accepting configurations* F' are those containing at least one accepting macrostate, i.e.,

$$F' = \{C \in P' \mid \exists M \in C \cdot \ M \text{ is an accepting macrostate}\}.$$

The transition relation above becomes deterministic on configurations: Let $\Delta'(M, a) = \{N_1, \ldots, N_k\}$ if $N_1, \ldots, N_k$ are *all* macrostates N s.t. $M \xrightarrow{a} N$

(there are finitely many such macrostates since M is a finite multiset), and for a configuration C, let

$$\Delta'(C, a) = \bigcup_{M \in C} \Delta'(M, a).$$

Then, we define the automaton $\mathcal{B}$ as $(\Sigma, P', I', F', \Delta')$.

Consider the following preorder[1] on configurations: $C \sqsubseteq^*_{\forall\exists} D$ if for every macrostate $M \in C$ there exists a macrostate $N \in D$ s.t. $M \sqsubseteq^* N$. Then,

$$C \sqsubseteq^*_{\forall\exists} D \wedge C \xrightarrow{a} C' \text{ implies } \exists D' \cdot C' \sqsubseteq^*_{\forall\exists} D' \wedge D \xrightarrow{a} D'.$$

Lemma 2. *The preorder $\sqsubseteq^*_{\forall\exists}$ is a wqo.*

Proof. By Higman's lemma we know that the preorder $\leqslant$ between finite multisets is a wqo, thus the same holds for the coarser $\sqsubseteq^*$, and consequently its lifting $\sqsubseteq^*_{\forall\exists}$ to finite subsets (a special case of finite multisets) is a wqo. □

We consider the following algorithm. Starting from the initial configuration C_I, we pick a configuration C and for every input symbol $a \in \Sigma$ s.t. $C \xrightarrow{a} D$ we add to the set the successor configuration D. If we reach a rejecting configuration, we conclude that the automaton is not universal. We employ the following two optimisations.

1. Configuration reduction: For every configuration C, we keep only those macrostates in C which are $\sqsubseteq^*$-maximal.
2. Configuration subsumption: If two configurations C, D exist s.t. $C \sqsubseteq^*_{\forall\exists} D$, then remove D.

[1] This is sometimes called the *Hoare preorder*, not to be confused with the *Smyth order* defined as $\forall N \in D \exists M \in C \cdot M \leqslant N$.