A NOTE ON FINITELY GENERATED SEMIGROUPS OF REGULAR LANGUAGES*

SERGEY AFONIN AND ELENA KHAZOVA
Lomonosov Moscow State University, Institute of Mechanics
Michurinskij pr., 1, Moscow, 119192, Russian Federation
E-mail: serg@msu.ru

Let $\mathcal{E} = \{E_1, \ldots, E_k\}$ be a set of regular languages over a finite alphabet Σ. Consider morphism $\varphi : \Delta^+ \to (\mathcal{S}, \cdot)$ where Δ^+ is the semigroup over a finite set Δ and $(\mathcal{S}, \cdot) = \langle \mathcal{E} \rangle$ is the finitely generated semigroup with $\mathcal{E}$ as the set of generators and language concatenation as a product. We prove that the membership problem of the semigroup $\mathcal{S}$, the set $[u] = \{v \in \Delta^+ \mid \varphi(v) = \varphi(u)\}$, is a regular language over Δ, while the set $\mathrm{Ker}(\varphi) = \{(u, v) \mid u, v \in \Delta^+ \ \varphi(u) = \varphi(v)\}$ need not to be regular. It is conjectured however that every semigroup of regular languages is automatic.

1. Introduction

Regular languages play an important role in both theoretical and practical aspects of computer science. As regular languages are closed under basic language operation it is interesting to ask whether one language may be represented in terms of other regular languages. The language factorization problem, i.e. the problem of representing a given regular language as a concatenation of other (regular) languages is a special case of such a representation. If the set of factors is fixed then language factorization may be considered as the membership problem for a finitely generated semigroup: a regular language $R \subseteq \Sigma^*$ belongs to the semigroup $\mathcal{S} = \langle E_1, \ldots, E_k \rangle$ if and only if there exists a sequence $i_1, i_2, \ldots, i_n$ of integers such that $1 \leqslant i_p \leqslant k$ $(p = 1 \ldots n)$ and

$$R = E_{i_1} E_{i_2} \ldots E_{i_n}.$$

The membership problem for a finitely generated semigroup of regular languages was shown to be decidable in [5]. This solution is based on reduc-

*This research was supported in part by the grant no. 10002-251 of the Russian Academy of Sciences Presidium program no. 17 "Parallel computations on multiprocessor systems"

tion to the so-called limitedness property of distance automata and gives an answer to the question whether a given regular language R belongs to $\mathcal{S}$ or not. Once we know that the answer is positive, a solution can be found by exhaustive search.

In [1] the authors considered finiteness conditions for semigroups of regular languages. It was found that a finitely generated semigroup $\mathcal{S} = \langle E_1, \ldots, E_k \rangle$ is finite if and only if for every set of non-repeating indices $\{i_1, \ldots, i_m\}$ $(m \leqslant k)$ there exists a natural number p such that $(E_{i_1} \ldots E_{i_m})^p = (E_{i_1} \ldots E_{i_m})^{p+1}$. In contrast, the semigroup given by the presentation $S = \langle \Delta \mid x^3 = x^2 \text{ for all } x \in \Delta^* \rangle$ is infinite [2] in the case of $|\Delta| \geqslant 2$. Actually, it is not surprising that semigroups of regular languages have a more complicated structure, because the structure of regular languages induces some additional relations between generators of a semigroup. In this paper we study the membership and word problems of the finitely generated semigroups of regular languages.

The layout of the paper is as follows. In section 2 we present basic definitions and briefly describe some useful results. In section 3 the regularity of the membership problem and the non-regularity of the word problem are proved. In the last section we discuss a connection between semigroups of regular languages and automatic semigroups.

2. Preliminaries

We assume familiarity with formal languages and finite automata, but recall, in order to fix the notation, the basic definitions. An *alphabet* is a finite non-empty set of *symbols*. A finite sequence of symbols from an alphabet Σ is called a *word* over Σ. The *empty* word is denoted by ε. A word $u = a_1 a_2 \ldots a_k$ is called a *scattered subword* of a word v (denoted as $u \sqsubseteq v$), if there exist $w_1, \ldots, w_{k+1} \in \Sigma^*$ such that $v = w_1 a_1 w_2 a_2 \ldots w_k a_k w_{k+1}$.

Any set of words is called a *language* over Σ. Σ^* denotes the set of all finite words (including the empty word), Σ^+ denotes the set of all nonempty words over Σ, $\varnothing$ is the empty language (containing no words), and 2^{Σ^*} is the set of all languages over Σ. The *union* of languages L_1 and L_2 is denoted by $L_1 + L_2$, concatenation by $L_1 L_2$, and iteration (or Kleene star) by L^*.

A *(non-deterministic) automaton* is a tuple $\mathcal{A} = \langle \Delta, Q, \rho, q_\circ, F \rangle$, where Q is a finite set of states, Δ is an input alphabet, $\rho \subseteq Q \times \Delta \times Q$ is a set of transitions, $q_\circ \in Q$ is the initial state and $F \subseteq Q$ is a set of final states. A successful path π in $\mathcal{A}$ is a sequence $(q_1, \delta_1, q_2)(q_2, \delta_2, q_3) \ldots (q_{m-1}, \delta_m, q_m)$

of transitions such that $q_1 = q_\circ$ and $q_m \in F$. We call the word $\delta_1\delta_2\ldots\delta_m$ the *label* of π. The language $L(\mathcal{A})$ of the automaton $\mathcal{A}$ is the set of all words w such that there exists a successful path π labeled by w. The language L is called *regular* if it is recognized by a finite automaton. The class of regular languages over Σ is denoted by $\mathrm{Reg}(\Sigma)$.

The set X of pairs of words (u, v), where $u = a_1 \ldots a_m$ and $v = b_1 \ldots b_m$ $(a_i, b_i \in \Delta)$ is called regular if there exists a finite automaton over $\Delta \times \Delta$ that recognizes X. This definition may be extended to the case when length of words u and v differs by introducing the *padding symbol* \$, and the mapping $\mu : \Delta^+ \times \Delta^+ \to (\Delta \cup \{\$\})^+ \times (\Delta \cup \{\$\})^+$ that appends the minimum number of padding symbols at the end of the shorter word in order to equalize their lengths.

Let Σ be an alphabet and $\mathcal{E} = \langle E_1, \ldots, E_k\rangle$ be a set of regular languages over Σ. The concatenation of regular languages is a regular language and we write $(\mathcal{S}, \cdot) = \langle \mathcal{E} \rangle$ for the finitely generated semigroup, generated by $\mathcal{E}$ with concatenation as a semigroup product. The homomorphism $\varphi : \Delta^+ \to (\mathrm{Reg}(\Sigma), \cdot)$ between the free semigroup Δ^+ and the semigroup of regular languages with concatenation is called a *regular language substitution.* With every semigroup $(\mathcal{S}, \cdot) = \langle E_1, \ldots, E_k\rangle$ of regular languages we can associate a language substitution $\varphi : \{\delta_1, \ldots, \delta_k\} \to \mathcal{S}$, defined by the rule $\delta_i \mapsto E_i$. Conversely, every regular language substitution $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ generates the semigroup $\mathcal{S}_\varphi = \langle\{\varphi(\delta) \mid \delta \in \Delta)\}\rangle$. For a regular language $L \subseteq \Delta^+$ by $\varphi(L)$ we mean the set $\varphi(L) = \{\varphi(w) \mid w \in L\}$ of regular languages over Σ.

Definition 2.1. Let $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ be a regular language substitution. The *maximal rewriting* of a regular language $R \subseteq \Sigma^*$ with respect to φ is the set

$$M_\varphi(R) = \{w \in \Delta^+ \mid \varphi(w) \subseteq R\}.$$

The following theorem is due to Calvanese *et al.* [3]

Theorem 2.1. *Let $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ be a regular language substitution. For any regular language $R \subseteq \Sigma^*$ the maximal rewriting $M_\varphi(R)$ is a regular language over Δ.*

3. Membership and word problems

The regularity of the membership problem for the semigroup of regular languages is a simple corollary from Theorem 2.1 and Higman's lemma, which may be stated as follows.

Lemma 3.1. *In every infinite sequence $\{u_i\}_{i\geqslant 1}$ of words over a finite alphabet there exist indices i and j, such that $u_i \sqsubseteq u_j$.*

Let $w = \delta_{i_1} \ldots \delta_{i_m}$ be a word over Δ and $A \subseteq \Delta$. We shall call the language $w \Uparrow A^* = A^*\delta_{i_1}A^*\delta_{i_2} \ldots A^*\delta_{i_m}A^*$ the *shuffle extension* of w. By $E(w, A)$ denote the language $(w \Uparrow A^*) \cap M_\varphi(\varphi(w))$. Clearly, $E(w, A)$ is a regular language for all $w \in \Delta^+$.

Proposition 3.1. *Let $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ be a regular language substitution, $u \in \Delta^+$, and $\Delta_0 = \{\delta \in \Delta \mid \varepsilon \in \varphi(\delta)\}$. For every $v \in E(u, \Delta_0)$ we have*

$$\varphi(u) = \varphi(v).$$

Proof. Let $\delta \in \Delta_0$. Consider the word $v = u_1\delta u_2$, where $u_1, u_2 \in \Delta^*$ and $u = u_1u_2$. We have

$$\varphi(u) \subseteq \varphi(v) \subseteq \varphi(u).$$

The first inclusion is due $\varepsilon \in \varphi(\delta)$ while the second one follows from the definition of the language $E(u, \Delta_0)$. □

Theorem 3.1. *Let $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ be a regular language substitution and w be a word in Δ^+. The membership problem for the semigroup $\mathcal{S}_\varphi$*

$$[w] = \{u \in \Delta^+ \mid \varphi(u) = \varphi(w)\}$$

is a regular language over Δ.

Proof. By Theorem 2.1 the set $M_\varphi(w) = \{u \in \Delta^+ \mid \varphi(u) \subseteq \varphi(w)\}$ is regular. Clearly $[w] \subseteq M_\varphi(w)$. Let $[w]$ be an infinite language. We prove now that there exists a finite subset $F \subseteq [w]$ satisfying

$$[w] = \bigcup_{u \in F} E(u, \Delta_0).$$

Note that $u \sqsubseteq v$ implies $E(v, \Delta_0) \subseteq E(u, \Delta_0)$, so without loss of generality we may assume that if the language F contains a word v then it does not contain subwords of v. The finiteness of F follows immediately from Lemma 3.1, and thus $[w]$ may be represented as a finite union of regular languages. □

This proof is not constructive because Higman's lemma does not provide an algorithm for the construction of the set F. In [1] the authors provide the algorithm that for a given regular language substitution $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$, a regular language $K \subseteq \Delta^+$, and a regular language $R \in \Sigma^*$ checks whether or not R belongs to the *rational set of regular languages* $\mathcal{R}(K, \varphi) = \{\varphi(w) \mid w \in K\}$. On the basis of this result the set F may be constructed by the following procedure (the input of this algorithm is a word $w \in \Delta^+$):

- Assign $K = \Delta^+$, $F = \varnothing$
- Repeat the following steps until $\varphi(w) \in \mathcal{R}(K, \varphi)$
  - Find the shortest word $v \in K$ such that $\varphi(v) = \varphi(w)$
  - Add v to F
  - Assign $K = K \setminus E(v, \Delta_0)$

Proposition 3.2. *The algorithm is correct.*

Proof.

(1) The algorithm always terminates by Lemma 3.1 — elements of the set F are not subwords of each other.

(2) On each round of the algorithm we have that F does not contain subwords of K.

(3) On each round $E(v, \Delta_0) \subseteq [w]$ by Proposition 3.1. □

The result of Theorem 3.1 is nonextensible in the following sense.

Theorem 3.2. *Let* $\varphi : \Delta^+ \to \mathrm{Reg}(\Sigma)$ *be a regular language substitution. The set*

$$\mathrm{Ker}(\varphi) = \{(u, v) \mid u, v \in \Delta^+ \ \varphi(u) = \varphi(v)\}$$

need not be regular.

Proof. We show that if the set $\mathrm{Ker}(\varphi)$ is regular then the equivalence problem for a rational set of regular languages, i.e. the problem to decide whether or not two given rational sets $\mathcal{R}_1 = (K_1, \varphi)$ and $\mathcal{R}_2 = (K_2, \varphi)$ are equal as sets of languages over Σ, is decidable. Then we reduce the later problem to the finite substitutions equivalence problem, that is known to be undecidable.

For a given regular language $K \in \Delta^*$ by $\overline{K}$ denote the closure of K with respect to φ:

$$\overline{K} = \varphi^{-1}(\varphi(K)) = \{u \in \Delta^+ \mid \exists v \in K \ \varphi(u) = \varphi(v)\}.$$

Let $\mathcal{R}_1 = (K_1, \varphi)$ and $\mathcal{R}_2 = (K_2, \varphi)$ be rational sets of regular languages. We have $\mathcal{R}_1 = \mathcal{R}_2$ if and only if $\overline{K_1} = \overline{K_2}$. Now, suppose that the set

$\mathrm{Ker}(\varphi)$ is regular, i.e. there exists a finite automaton M that recognizes this language. By standard direct product construction of automata M and K we construct the automaton that recognizes the language $\{v \in \Delta^+ \mid \exists u \in K\ (u, v) \in \mathrm{Ker}(\varphi)\}$. Thus, if $\mathrm{Ker}(\varphi)$ is regular then the so is $\overline{K}$ for every regular language $K \in \Delta^*$.

Let φ_1 and φ_2 be finite substitutions, i.e. homomorphisms between Δ^+ and a semigroup of finite languages. The equivalence problem of finite substitutions on a regular language L

$$\varphi_1(w) = \varphi(w) \text{ for all } w \in L$$

is known to be undecidable [6] for $L = xy^*z$. Let $\Delta = \{x_1, y_1, z_1, x_2, y_2, z_2\}$, φ be a finite substitution, and rational sets $\mathcal{R}_1 = (K_1, \varphi)$ and $\mathcal{R}_2 = (K_2, \varphi)$ are given by languages $K_1 = x_1y_1^*z_1$ and $K_2 = x_2y_2^*z_2$. By considering the length of the longest word in the image $\varphi(w)$ we have that $\mathcal{R}_1$ and $\mathcal{R}_2$ are equal if and only if finite substitutions φ_1 and φ_2 (induced by φ) are equal on the language xy^*z. We have a contradiction, so the set $\mathrm{Ker}(\varphi)$ is not regular in general. □

4. Connection with automatic semigroups

Let us recall the definition of automatic semigroups [4]. Let S be a semigroup, A be a finite set, L be a regular language over A, and $\psi : A^+ \to S$ be a homomorphism with $\psi(L) = S$. The pair (A, L) is called an *automatic structure* for S if

(1) $L_= = \{(u, v) \mid u, v \in L, \psi(u) = \psi(v)\}$ is regular;
(2) $L_a = \{(u, v) \mid u, v \in L, \psi(ua) = \psi(v)\}$ is regular for each $a \in A$.

If a semigroup S has an automatic structure (A, L) for some A and L, then S is called *automatic.*

Automatic semigroups include many naturally appearing semigroups, e.g. finite semigroups and finitely generated subsemigroups of a free semigroup are automatic [4]. An attractive property of automatic semigroups is the solvability of the word problem in quadratic time.

Example 4.1. Let the semigroup $\mathcal{S}$ be generated by languages

$$x = (a + b)^*a, \quad y = \varepsilon + a + b, \text{ and } z = b^*$$

over $\Sigma = \{a, b\}$. In order to show that $\mathcal{S}$ is automatic we find the presentation for $\mathcal{S}$ and construct automata for $L_=$ and L_a.

First, the language z is a star, so $z^k = z^p$ for all $k, p \geqslant 1$. Second, since x is the set of all words over Σ that are ended by the letter a and both y and z contain the empty word we have equations $(y+z)^k x = x$. Finally, the language $xz = (a+b)^* ab^*$ is the set of all words over Σ that contain at least one letter a, so we have relations $x(y+z)^k z(y+z)^p = xz$ for all $k, p \geqslant 0$. The semigroup $\mathcal{S}$ satisfies no other relations. Thus, the semigroup $\mathcal{S}$ is given by the presentation

$$\langle x, y, z \mid yx = x, zx = x, z^2 = z, xzy = xz, xy^k z = xz \ (k \geqslant 1) \rangle.$$

We show now that $\mathcal{S}$ is automatic. First, let us construct the set L of normal forms, such that each element of the semigroup is represented by only one word in L. If a word $w \in L$ contains x then x is the first letter of the word. If w contains both x and z then w has the from $x^k z$ by the last three relations. If $w \in L$ does not contain x, then w is the set of all words over $\{y, z\}$ that does not contain zz as a subword. So we have

$$L = xx^*(y^* + z) + yy^*(zyy^*)^*(\varepsilon + z) + z(yy^*z)^*y^*.$$

If every element of a semigroup is represented by exactly one word in L then the set $L_= = \{(u, u) \mid u \in L\}$ is regular.

Now consider the multiplication by generators. The product of a word $u \in L$ by x equals to x if the word u does not contain x, and equals to x^{k+1} if $u = x^k u'$ (the word u' does not contain x). We have

$$\begin{aligned} L_x = {} & (x,x)(x,x)^* \, [(\$,x) + (y,x)(y,\$)^* + (z,x)] + \\ & (y,x)(y,\$)^*[(z,\$)(y,\$)(y,\$)^*]^*[\varepsilon + (z,\$)] + \\ & (z,x) \, [(y,\$)(y,\$)^*(z,\$)]^* \, (y,\$)^*. \end{aligned}$$

The first line above corresponds to the first part of $L_=$, i.e. words that start with x, the second and the third lines – to words that start with y and z, respectively.

Similarly we can construct the languages L_y and L_z. We have

$$\begin{aligned} L_y = {} & (x,x)(x,x)^*[(y,y)^*(\$,y) + (z,z)] + \\ & (y,y)(y,y)^*[(z,z)(y,y)(y,y)^*][\varepsilon + (z,z)](\$,y) + \\ & (z,z)[(y,y)(y,y)^*(z,z)]^*(y,y)^*(\$,y) \end{aligned}$$

and

$$\begin{aligned} L_z = {} & (x,x)(x,x)^*[(\$,z) + (y,z)(y,\$)^* + (z,z)] + \\ & (y,y)(y,y)^*[(z,z)(y,y)(y,y)^*][(\$,z) + (z,z)] + \\ & (z,z)[(y,y)(y,y)^*(z,z)]^*[(y,y)(y,y)^*(\$,z) + (z,\$)]. \end{aligned}$$

The language $L_=$ and all languages L_a for $a \in \Delta$ are regular so the semigroup $\mathcal{S}$ is automatic.

Theorem 3.2 states that in general (A, Δ^+) is not an automatic structure for a semigroup of regular languages. Nevertheless, we expect that

Conjecture 4.1. *Every semigroup of regular languages is automatic.*

Although automatic structure may be easily constructed for a particular semigroup of regular languages, just like in the above example, this conjecture seems not to be trivial. For example, let $\mathcal{S}$ be a semigroup of finite languages. Every finite language may be factorized into prime languages, but this factorization is not unique (e.g. $\{\varepsilon, a, a^2, a^3\} = \{\varepsilon, a\}^3 = \{\varepsilon, a\}\{\varepsilon, a^2\}$), thus prime factors may satisfy nontrivial relations. In the general case the situation become more complicated.

References

1. S. Afonin and E. Hazova, Membership and finiteness problems for rational sets of regular languages, Proceedings of the Developments in Language Theory 2005, C. De Felice and A. Restivo, Eds., *Lecture Notes in Computer Science* **3572**, Springer, 88–99 (2005).
2. J. Brzozowski, K. Culik, and A. Gabrielian, Classification of noncounting events, *Journal of Computer and System Sciences* **5**, 41–53 (1971).
3. D. Calvanese, G. De Giacomo, M. Lenzerini and M. Vardi, Rewriting of regular expressions and regular path queries, *Journal of Computer and System Sciences* **64**, 443–465 (2002).
4. C. M. Campbell, E. F. Robertson, N. Ruškuc and R. M. Thomas, Automatic semigroups, *Theoretical Computer Science* **250 (1–2)**, 365–391 (2001).
5. K. Hashiguchi, Representation theorems on regular languages, *Journal of Computer and System Sciences* **27**, 101–115 (1983).
6. J. Karhumäki and L. P. Lisovik, The equivalence problem of finite substitutions on ab*c, with applications, *ICALP'02: Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, Springer-Verlag, 812–820 (2002).