Journal of Universal Computer Science, vol. 8, no. 2 (2002), 257-259
submitted: 15/9/02, accepted: 27/2/02, appeared: 28/2/02 © J.UCS

# Shuffle Decomposition of Regular Languages[1]

Masami Ito
(Department of Mathematics, Kyoto Sangyo University
Kyoto 603-8555, Japan
Email: ito@ksuvx0.kyoto-su.ac.jp)

**Abstract:** Let $A \subseteq X^*$ be a regular language. In the paper, we will provide an algorithm to decide whether there exist a nontrivial language $B \in \mathcal{I}(n, X)$ and a nontrivial regular language $C \subseteq X^*$ such that $A = B \diamond C$

**Key Words:** regular language, shuffle product, shuffle decomposition, $\mathcal{I}(n, X)$
**Category:** F.1

In this paper, we will deal with shuffle decompositions of regular languages over an alphabet $X$. Regarding definitions and notations concerning formal languages and automata, not defined in this paper, refer, for instance, to [1]. Now let $\mathcal{A} = (S, X, \delta, s_0, F)$ be a finite automaton with $\mathcal{L}(\mathcal{A}) = A$ and let $\mathcal{B} = (T, X, \gamma, t_0, G)$ be a finite automaton with $\mathcal{L}(\mathcal{B}) = B$. We will look for a regular language $C$ over $X$ such that $A = B \diamond C$. By $\overline{X}$, we denote the language $\{\overline{a} \mid a \in X\}$ with $X \cap \overline{X} = \emptyset$. Let $\overline{\mathcal{B}} = (T, X \cup \overline{X} \cup \{\#\}, \overline{\gamma}, t_0, G)$ where $\overline{\gamma}$ is defined as follows:

For $t \in T$ and $a \in X$, $\overline{\gamma}(t, a) = t$, $\overline{\gamma}(t, \overline{a}) = \gamma(t, a)$. Moreover, $\overline{\gamma}(t, \#) = t$ if $t \in G$.

Then the following can be easily shown.

**Fact 1** *Let* $a_1 a_2 \ldots a_n \in X^*$ *where* $a_i \in X, i = 1, 2, \ldots, n$. *Then* $a_1 a_2 \ldots a_n \in \mathcal{L}(\mathcal{B})$ *if and only if* $u_1 \overline{a}_1 u_2 \overline{a}_2 \ldots u_n \overline{a}_n u_{n+1} \# \in \mathcal{L}(\overline{\mathcal{B}})$ *where* $u_1, u_2, \ldots, u_n \in X^*$.

Let $\mathcal{A}_1 = (\overline{S}, X \cup \overline{X} \cup \{\#\}, \overline{\delta}, s_0, \{\alpha, \omega\})$ and let $\mathcal{A}_2 = (\overline{S}, X \cup \overline{X} \cup \{\#\}, \overline{\delta}, s_0, \{\alpha\})$ where $\overline{S} = (\cup_{a \in X \cup \{\lambda\}} S^{(a)}) \cup \{\alpha, \omega\}$. Here $S^{(\lambda)}$ is regarded as $S$ where $\lambda$ is the empty word. For $s \in S, t \in S \setminus F, t' \in F, a \in X \cup \{\lambda\}, b \in X$ and $\{\#\}$, $\overline{\delta}$ is defined as follows:

$\overline{\delta}(s^{(a)}, b) = \delta(s, b)^{(a)}, \overline{\delta}(s^{(a)}, \overline{b}) = \delta(s, b)^{(b)}, \overline{\delta}(t^{(a)}, \#) = \{\alpha\}$ and $\overline{\delta}(t'^{(a)}, \#) = \{\omega\}$.

We consider the following two automata:

$\mathcal{C}_1 = (\overline{S} \times T, X \cup \overline{X} \cup \{\#\}, \overline{\delta} \times \overline{\gamma}, (s_0, t_0), \{\alpha, \omega\} \times G)$, $\mathcal{C}_2 = (\overline{S} \times T, X \cup \overline{X} \cup \{\#\}, \overline{\delta} \times \overline{\gamma}, (s_0, t_0), \{\alpha\} \times G)$ where $\overline{\delta} \times \overline{\gamma}((\bar{s}, t), a) = (\overline{\delta}(\bar{s}, a), \overline{\gamma}(t, a))$ for $(\bar{s}, t) \in \bar{S} \times T$ and $a \in X$.

Now consider the following homomorphism $\rho$ of $(X \cup \overline{X} \cup \{\#\})^*$ into $X^*$:

$\rho(a) = a$ for $a \in X$, $\rho(\overline{a}) = \lambda$ for $a \in X$ and $\rho(\#) = \lambda$.

[1] C. S. Calude, K. Salomaa, S. Yu (eds.). *Advances and Trends in Automata and Formal Languages. A Collection of Papers in Honour of the 60th Birthday of Helmut Jürgensen.*

**Lemma 1.** *Automata accepting the languages $\rho(\mathcal{L}(\mathcal{C}_1))$ and $\rho(\mathcal{L}(\mathcal{C}_2))$ can be effectively constructed.*

*Proof.* Let $i = 1, 2$. From $\mathcal{C}_i$, we can construct a regular grammar $\mathcal{G}_i$ such that $\mathcal{L}(\mathcal{G}_i) = \mathcal{L}(\mathcal{C}_i)$ with the production rules of the form $A \to aB$ ($A, B$ are variables and $a \in X \cup \overline{X} \cup \{\#\}$). Replacing every rule of the form $A \to aB$ in $\mathcal{G}_i$ by $A \to \rho(a)B$, we can obtain a new grammar $\mathcal{G}_i'$. Then it is clear that $\rho(\mathcal{L}(\mathcal{C}_i)) = \mathcal{L}(\mathcal{G}_i')$. Using this grammar $\mathcal{G}_i'$, we can construct an automaton $\mathcal{D}_i$ with $\lambda$-move such that $\mathcal{L}(\mathcal{D}_i) = \mathcal{L}(\mathcal{G}_i')$ i.e. $\rho(\mathcal{L}(\mathcal{C}_i)) = \mathcal{L}(\mathcal{D}_i)$. Notice that all the above procedures are effectively done. This completes the proof of the lemma.

Let $B, C \subseteq X^*$. By $B \diamond C$ we denote the shuffle product of $B$ and $C$, i.e. $\{u_1v_1u_2v_2 \dots u_nv_n \mid u = u_1u_2 \dots u_n \in B, v = v_1v_2 \dots v_n \in A\}$.

**Proposition 2.** *Let $u \in X^*$. Then $\{u\} \diamond B \subseteq A$ if and only if $u \in \rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$.*

*Proof.* ($\Rightarrow$) Let $u = u_1u_2 \dots u_nu_{n+1} \in X^*$ and let $a_1a_2 \dots a_n \in B$ where $u_1, u_2, \dots, u_n, u_{n+1} \in X^*$ and $a_1, a_2, \dots, a_n \in X$. Then $\overline{\delta} \times \overline{\gamma}((s_0, t_0), u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) = (\overline{\delta}(s_0, u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#), \overline{\gamma}(t_0, u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#))$ $= (\overline{\delta}(\delta(s_0, u_1a_1u_2a_2 \dots u_na_nu_{n+1})^{(a_n)}, \#), \overline{\gamma}(\gamma(s_0, a_1a_2 \dots a_n)^{(a_n)}, \#)) = (\omega, \gamma$ $(t_0, a_1a_2 \dots a_n)) \in \{\omega\} \times G$. Therefore, $u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\# \in \mathcal{L}(\mathcal{C}_1) \setminus \mathcal{L}(\mathcal{C}_2)$. Hence $u = u_1u_2 \dots u_nu_{n+1} = \rho(u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) \in \rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$. ($\Leftarrow$) Suppose that $u \diamond B \subseteq A$ does not hold though $u \in \rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$. Then there exist $u = u_1u_2 \dots u_nu_{n+1} \in X^*$ and $a_1a_2 \dots a_n \in B$ such that $u_1a_1u_2a_2 \dots u_na_nu_{n+1} \notin A$. Hence $\overline{\gamma}(t_0, u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) = \overline{\gamma}(\gamma(t_0, a_1a_2 \dots a_n), \#) = \gamma(t_0, a_1a_2 \dots a_n) \in G$. On the other hand, since $u_1a_1u_2a_2 \dots u_na_n u_{n+1} \notin A$, $\overline{\delta}(s_0, u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) = \overline{\delta}(\delta(s_0, u_1a_1u_2a_2 \dots u_na_nu_{n+1})^{(a_n)}, \#) = \{\alpha\}$. Hence $\overline{\delta} \times \overline{\gamma}((s_0, t_0), u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) \in \{\alpha\} \times G$, i.e. $u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\# \in \mathcal{L}(\mathcal{C}_2)$. Therefore, $u = \rho(u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\#) \in \rho(\mathcal{L}(\mathcal{C}_2))$. On the other hand, it is obvious that $u_1\overline{a}_1u_2\overline{a}_2 \dots u_n\overline{a}_nu_{n+1}\# \in \mathcal{L}(\mathcal{C}_1)$. Thus $u \notin \rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$, a contradiction. Consequently, the proposition must hold true.

**Corollary 3.** *In the above, $B \diamond (\rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))) \subseteq A$.*

Let $L \subseteq X^*$ be a regular language over $X$. By $\#L$, we denote the number $min\{|S| \mid \exists \mathcal{A} = (S, X, \delta, s_0, F), L = \mathcal{L}(\mathcal{A})\}$ where $|S|$ denotes the cardinality of $S$. Moreover, $\mathcal{I}(n, X)$ denotes the class of languages $\{L \subseteq X^* \mid \#L \leq n\}$.

**Theorem 4.** *Let $A \subseteq X^*$ and let $n$ be a positive integer. Then it is decidable whether there exist nontrivial regular languages $B \in \mathcal{I}(n, X)$ and $C \subseteq X^*$ such that $A = B \diamond C$. Here a language $D \subseteq X^*$ is said to be nontrivial if $D \neq \{\lambda\}$.*

*Proof.* Let $A \subseteq X^*$ be a regular language. Assume that there exist nontrivial regular languages $B \in \mathcal{I}(n, X)$ and $C \subseteq X^*$ such that $A = B \diamond C$. Then, by Proposition 2 and its corollary, $C \subseteq \rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$ and $B \diamond (\rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))) \subseteq A$. Hence $A = B \diamond (\rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2)))$. Thus we have the following algorithm: (1) Choose a nontrivial regular language $B \subseteq X^*$ from $\mathcal{I}(n, X)$ and construct the language $\rho(\mathcal{L}(\mathcal{C}_1)) \setminus \rho(\mathcal{L}(\mathcal{C}_2))$ (see Lemma 1). (2) Let $C = \rho(\mathcal{L}(\mathcal{C}_1)) \setminus$

$\rho(\mathcal{L}(\mathcal{C}_2))$. (3) Compute $B \diamond C$. (4) If $A = B \diamond C$, then the output is "YES" and "NO", otherwise. (4) If the output is "NO", then choose another element in $\mathcal{I}(n, X)$ as $B$ and continue the procedures (1) - (3). (5) Since $\mathcal{I}(n, X)$ is a finite set, the above process terminates after a finite-step trial. Once one gets the output "YES", then there exist nontrivial regular languages $B \in \mathcal{I}(n, X)$ and $C \subseteq X^*$ such that $A = B \diamond C$. Otherwise, there are no such languages.

Let $n$ be a positive integer. By $\mathcal{F}(n, X)$, we denote the class of finite languages $\{L \subseteq X^* \mid max\{|u| \mid u \in L\} \leq n\}$ where $|u|$ is the length of $u$. Then the following result by C. Câmpeanu et al. ([2]) can be obtained as a corollary of Theorem 4.

**Corollary 5.** *For a given positive integer $n$ and a regular language $A \subseteq X^*$, the problem whether $A = B \diamond C$ for a nontrivial language $B \in \mathcal{F}(n, X)$ and a nontrivial regular language $C \subseteq X^*$ is decidable.*

*Proof.* Obvious from the fact that $\mathcal{F}(n, X) \subseteq \mathcal{I}(|X|^{n+1}, X)$.

**Acknowledgement** The author would like to thank Dr. Kai Salomaa for his usuful comments.

## References

1. J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading MA,1979.
2. C. Câmpeanu, K. Salomaa, S. Vágvölgyi, Shuffle quotient and decompositions, Lecture Notes in Computer Science (Springer), to appear.