A-010

Morphic Characterizations with Insertion and Locality in the Framework of Chomsky-Schützenberger Theorem

藤岡　薫 †
Kaoru Fujioka

1. Introduction

Representing a class of languages through operations on its subclasses is a traditional issue within formal language theory. For context-free languages, there is a well-known Chomsky-Schützenberger characterization using its subclasses: each context-free language L can be represented in the form $L = h(D \cap R)$, where D is a Dyck language (parenthesis language), R is a regular language, and h is a projection [1].

Among the variety of representation theorems for context-free languages [7] [5], Chomsky-Schützenberger theorem is unique in that it consists of Dyck languages, regular languages, and simple operations. In this work, we obtain some characterizations and representation theorems of context-free languages and regular languages in Chomsky hierarchy by insertion systems, strictly locally testable languages, and morphisms in the framework of Chomsky-Schützenberger theorem.

2. Preliminaries

An *insertion system* is a triple $\gamma = (T, P, A_X)$, where T is an alphabet, P is a finite set of *insertion rules* of the form (u, x, v) with $u, x, v \in T^*$, and A_X is a finite set of strings over T called axioms. We write $\alpha \overset{r}{\Longrightarrow}_\gamma \beta$ if $\alpha = \alpha_1 uv\alpha_2$ and $\beta = \alpha_1 uxv\alpha_2$ for some insertion rule $r : (u, x, v) \in P$ with $\alpha_1, \alpha_2 \in T^*$. A language generated by γ is defined by $L(\gamma) = \{w \in T^* \mid s \Longrightarrow_\gamma^* w \text{ for some } s \in A_X\}$.

An insertion system γ is said to be of *weight* (i, j) if $i = max\{|x| \mid (u, x, v) \in P\}$, $j = max\{|u| \mid (u, x, v) \in P \text{ or } (v, x, u) \in P\}$. For $i, j \geq 0$, INS_i^j is the class of all languages generated by insertion systems of weight (i', j') with $i' \leq i$ and $j' \leq j$.

From the definition, for any $0 \leq i' \leq i$ and $0 \leq j' \leq j$, we have an inclusion $INS_{i'}^{j'} \subseteq INS_j^i$ [6].

Furthermore, it has already been known that for any $i \geq 1$, INS_i^1 is properly included by the class of context-free languages, denoted by CF [6] and INS_1^0 is properly included by the class of regular languages, denoted by REG [2].

For a positive integer $k \geq 1$, a language L over an alphabet T is *strictly k-testable* if there is a triplet (A, B, C) with $A, B, C \subseteq T^k$ such that for any w with $|w| \geq k$, w is in L iff the prefix of w of length k is in A, the suffix of w of length k is in B, and any proper interior substring of w of length k is in C [3]. Let $LOC(k)$ be the class of strictly k-testable languages.

It has already been known the proper inclusion $LOC(1) \subset LOC(2) \subset \cdots \subset LOC(k) \subset \cdots \subset REG$, where REG is the class of regular languages [4].

For languages L_1, L_2, and a morphism h, we use the following notation: $h(L_1 \cap L_2) = \{h(w) \mid w \in L_1 \cap L_2\}$. For language classes $\mathcal{L}_1$ and $\mathcal{L}_2$, we introduce the following language class:

$$H(\mathcal{L}_1 \cap \mathcal{L}_2) = \{h(L_1 \cap L_2) \mid \begin{array}{l} h \text{ is a morphism,} \\ L_i \in \mathcal{L}_i \ (i = 1, 2)\}. \end{array}$$

We show characterizations and representation theorems of language families in Chomsky hierarchy with this notation.

3. Characterizations by context-free insertion systems

In this section, using insertion systems of weight $(i, 0)$ for $i \geq 1$, in which no insertion operation can be controlled by contexts, we prove the following characterizations concerning the class of regular languages and the class of context-free languages.

We prove the representation theorem for regular languages with the help of strictly 2-testable languages such that $REG = H(INS_1^0 \cap LOC(2))$. To show this, we consider the inclusion $REG \supseteq H(INS_1^0 \cap LOC(2))$, which can be derived from $INS_1^0 \subset REG$, $LOC(2) \subset REG$, and the closure properties of regular languages. The other inclusion is proved by simulating the derivations in regular grammar using insertion systems and strictly 2-testable languages by the induction.

In contrast to this, we show that a regular language $L = \{a^l \mid l \geq 0\} \cup \{b^l \mid l \geq 0\}$ cannot be written in the form $h(L(\gamma) \cap R)$, for any insertion system γ of weight $(i, 0)$ $(\forall i \geq 1)$, strictly 1-testable language R, and morphism h by contradiction. Then we have the proper inclusion $H(INS_1^0 \cap LOC(1)) \subset REG$.

The value of weight $(1, 0)$ in insertion systems is optimal for expressing regular languages which is shown by considering the following example: For an insertion system $\gamma = (\{a, b\}, \{(\lambda, ab, \lambda)\}, \{\lambda\})$, a strictly 1-testable language R prescribed by $A = B = C = \{a, b\}$, and a morphism $h : T^* \to T^*$ such as $h(c) = c$ $(\forall c \in T)$, it is shown that $L(\gamma) \cap R = h(L(\gamma) \cap R) =$

†Office for Strategic Research Planning, Kyushu University, 6-10-1 Hakozaki Higashi-ku Fukuoka-shi, 812-8581, Japan.
kaoru@tcslab.csce.kyushu-u.ac.jp

$\{w \mid w \in L(\gamma),\ w \neq \lambda\}$ is not regular by considering a language $L(\gamma) \cap R \cap \{a^i b^j \mid i, j \geq 1\} = \{a^i b^i \mid i \geq 1\}$, which is not regular.

For context-free languages, we prove the representation theorem such as $CF = H(INS_2^0 \cap LOC(2))$. To show this, the inclusion $CF \subseteq H(INS_2^0 \cap LOC(2))$ is proved by simulating a derivation of context-free grammar in Chomsky normal form. The other one is proved by using $INS_2^0 \subset CF$, $LOC(2) \subset REG$, and closure properties of context-free languages.

4. Characterizations by insertion systems with a context of length one

Now we consider the characterization and representation theorems using insertion systems of weight $(i, 1)$ with $i \geq 1$. As is shown in [6], INS_i^1 is known to be a proper subset of the class of context-free languages. We prove that INS_i^1 with $i \geq 1$ is incomparable with the class of regular languages by considering an insertion system $\gamma = (\{a, b, c, d\}, \{(a, c, b), (c, d, b), (c, a, d), (a, b, d)\}, \{ab\})$ and a regular language $L = \{a^{2in} \mid n \geq 1\}$.

With the help of strictly 1-testable languages and simple operations, we show the inclusion $INS_i^1 \subseteq H(INS_i^1 \cap LOC(1))$. The inclusion can be derived easily if we consider the fact $V^* \in LOC(1)$ for any alphabet V.

Furthermore, we show a characterization of context-free languages using insertion systems of weight $(i, 1)$ with $i \geq 1$ and strictly 1-testable languages. The inclusion $H(INS_i^1 \cap LOC(1)) \subseteq CF$ $(i \geq 1)$ can be derived directly from the fact $INS_i^1 \subset CF$, $LOC(1) \subset REG$, and the closure property of context-free languages. We prove the proper inclusion $H(INS_i^1 \cap LOC(1)) \subset CF$ $(i \geq 1)$ by showing that for a context-free language $L = \{a^n b a^n \mid n \geq 1\}$, there are no insertion system γ of weight $(i, 1)$, strictly 1-testable language R, and morphism h such that $L = h(L(\gamma) \cap R)$ by contradiction.

In contrast to this, we prove the representation theorem for the class of context-free languages by insertion systems of weight $(1, 1)$ with the help of strictly 2-testable languages, that is, $H(INS_1^1 \cap LOC(2)) = CF$.

To show the representation theorem, from Chomsky-Schützenberger theorem, $CF = H(Dyck \cap REG)$, we consider the inclusion $H(INS_1^1 \cap LOC(2)) \supseteq H(Dyck \cap REG)$, where $Dyck$ is the class of Dyck languages. For any language $L = h(D \cap L(G))$ with Dyck language D, regular grammar G, and morphism h, we construct an insertion system γ of weight $(1, 1)$, a strictly 2-testable language R, and a morphism h' and prove that $h'(L(\gamma) \cap R) = L$ holds.

The converse inclusion $H(INS_1^1 \cap LOC(2)) \subseteq CF$ can be derived directly from the fact that $INS_1^1 \subset CF$, $LOC(2) \subset REG$, and the closure property of context-free languages.

5. Conclusion

In this work, we contribute to the study of insertion systems controlled by a context of length 1 and context-free insertion systems for new characterizations of context-free and regular languages.

We showed the following:

- $H(INS_1^0 \cap LOC(k))) = REG$ with $k \geq 2$.
- $H(INS_1^0 \cap LOC(1)) \subset REG \subset H(INS_i^0 \cap LOC(k))$ with $i, k \geq 2$.
- $H(INS_i^0 \cap LOC(k)) = CF$ with $i, k \geq 2$.
- $H(INS_i^1 \cap LOC(1)) \subset CF$ with $i \geq 1$.
- $H(INS_i^1 \cap LOC(k)) = CF$ with $i \geq 1, k \geq 2$.

Acknowledgments

This work was supported in part by Grants-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 23740081).

References

[1] Chomsky, N., Schutzenberger,M.P. The algebraic theory of context-free languages. *Computer Programming and Formal Systems*, pp.118-161, 1963.

[2] Fujioka, K. Morphic characterizations of languages in Chomsky hierarchy with insertion and locality. *Inf. Comput.*, **209**, **3**, pp.397-408, 2011.

[3] Head, T. Splicing representations of strictly locally testable languages *Discrete Applied Math*, **87**, pp.87-139, 1998.

[4] McNaughton,R., Papert,S.A. Counter-Free Automata (M.I.T. research monograph no. 65). *The MIT Press*, 1971.

[5] Okubo, F., Yokomori, Y. Morphic Characterizations of Language Families in Terms of Insertion Systems and Star Languages. *Int. J. Found. Comput. Sci.*, **22**, **1**, pp. 247-260, 2011.

[6] Păun, G., Rozenberg, G., Salomaa, A. DNA Computing. New Computing Paradigms. *Springer*, 1998.

[7] Rozenberg, G., Salomaa, A. Handbook of formal languages *Springer-Verlag New York, Inc.*, 1997.