第3卷 第1期 1983年 1月
数学研究与评论
JOURNAL OF MATHEMATICAL RESEARCH AND EXPOSITION
Vol.3 No.1 Jan. 1983

The Improvement in Chomsky-Schützenberger Theorem and Its Form in "Deterministic" Case*

Li Lian (李 廉)

(The department of Math. and Mech. Lanzhou University)

In several papers such as [1], [2], [4], [8], the representation problems of formal languages were discussed. It is well-known, Chomsky-Schützenberger theorem plays an important role in the representation of context-free languages which is one of the most interesting classes of languages. The Chomsky-Schützenberger theorem asserts that given a Σ_1, there exist Σ_2, a Dyck set $D_{\Sigma_2} \subseteq \hat{\Sigma}_2^*$ and a homomorphism h from $\hat{\Sigma}_2^*$ onto Σ_1^* which satisfy the property that for each context-free language $L \subseteq \Sigma_1^*$ a regular set $R \subseteq \hat{\Sigma}_2^*$ can be found such that $h(D_{\Sigma_2} \cap R) = L$.

In present paper, we establish a normal-form of pushdown automata, abbreviated pda. and improve the Chomsky-Schützenberger theorem. From this improved theorem it follows that the context-free languages form a principal AFL without use of AFA[2]. We give a variant of this improved theorem in "deterministic" case. As a co-product we obtain a necessary condition for a deterministic context-free language can be accepted in real-time by a deterministic pushdown automaton, abbreviated dpda, with empty store, this result implies that it is not all of $LR(k)$ grammars can be parsed in real-time.

1. Preliminary

Definition 1 A language L on $\sum$ is called a root language if there exists a set B such that $L = B^*$ and the B is called a root of L[3],[4],[5]. A root language L on $\sum$ is called bi-simple-root language, abbreviated BSR language, if the smallest root of L is both prefix and suffix set.

Definition 2 Let Σ_1, Σ_2 be alphabets, $L \subseteq \Sigma_1^*$, a homomorphism h: $\Sigma_1^* \to \Sigma_2^*$ is called k-limited for L, where k is a nonnegative integer, if for any $xuy \in L$ and $h(u) = \varepsilon$ follows $|u| \leqslant k$. If L is understood, we call that h is k-limited or limited.

*Received Sept. 29, 1981

Definition 3 Let Σ_1, Σ_2 be alphabets, $L\subseteq\Sigma_1^*$, a homomotrphism h: $\Sigma_1^*\to\Sigma_2^*$ is called essential injective for L, if x, $y\in L$ and $f(x)=f(y)$ follows $x=y$. If L is understood, we call that h is essential injective.

Let Σ be a alphabet, denote $\overline{\Sigma}=\{\overline{a}\mid a\in\Sigma\}$, and $\hat{\Sigma}=\Sigma\cup\overline{\Sigma}$. For a pda. we denote by $N(M)$ the language accepted by M with empty store and by $T(M)$ the language accepted by M with final states.

Theorem 1 (**Normal-form of pda.**) For any pda. which does not accept ε, there exist pda's M' and M'' with no ε-move satisfying

(*) if $(p, \alpha)\in\delta(q, a, A)$, then $\alpha=BA$ or $\alpha=A$ or $\alpha=\varepsilon$,

where $B\in\Gamma$, such that $L(M)=N(M')=T(M'')$.

Remark 1 From Theorem 1, it is easy to see that for any (real-time) dpda there exists a (real-time) dpda M' satisfying (*) such that $N(M)=N(M')$ $(T(M)=T(M'))$.

2. The improvement in the Chomsky-Schützenberger theorem

Lemma 1 For any context-free language L on Σ, there exists a grammar $G=(V_N, \Sigma, P, S,)$ satisfying

(i) Every production is of the form $S\to\varepsilon$ or $A\to a$ or $A\to aB$ or $A\to aBC$, where $a\in\Sigma$ and A, B $C\in V_N$;

(ii) S never occur in the right of any production, such that $L(G)=L$.

Lemma 1 was stated in [6], but we may provide a new proof, although we omit it herf.

Theorem 2 (Improvement in Chomsky-Schützenberger theorem) For any given alphabet Σ_1, there exist Σ_2, a Dyck set $D_{\Sigma_2}\subseteq\hat{\Sigma}_2^*$, a word $w\in\hat{\Sigma}_2^*$ and a homomorphism h from $\hat{\Sigma}_2^*$ onto Σ_1^* which satisfy the property that for each context-free language $L\subseteq\Sigma_1^*$, a regular BSR language $L_R\subseteq\hat{\Sigma}_2^*$ can be found such that $L=h(D_{\Sigma_2}\cap wL_R)$, and h is limited for $D_{\Sigma_2}\cap wL_R$.

The outline of proof Let Σ_1 be a given alphabet, let $\Sigma_2=\Sigma_1\cup\{c, d\}$, c, $d\notin\Sigma_1$ and h: $\hat{\Sigma}_2^*\to\Sigma_1^*$ be a homomorphism defined as follows: $h(a)=\varepsilon$, $\forall a\in\overline{\Sigma}_2\cup\{c, d\}$, and $h(a)=a$, $\forall a\in\Sigma_1$. From Lemma 1, for any context-free language $L\subseteq\Sigma_1^*$, there exists a grammar G satisfying the conditions in Lemma 1 such that $L=L(G)$. Let

$$G=(\{X_1, X_2,\cdots, X_n\}, \Sigma, P, X_1)\quad P=\{\pi_1, \pi_2, \cdots, \pi_m,\}.$$

Define a homomorphism g: $P^*\to\hat{\Sigma}_2^*$ as follows

(i) if $\pi_i = X_j \to aX_{j_1}X_{j_2}$, then $g(\pi_i) = a\bar{a}\,\bar{c}\,\overline{d^j}\,\bar{c}\,cd^{j_1}ccd^{j_2}c$;

(ii) if $\pi_i = X_j \to aX_{j_1}$, then $g(\pi_i) = a\bar{a}\,c\,\bar{c}\,\bar{c}\,\overline{d^j}\,\bar{c}\,cd^{j_1}c$;

(iii) if $\pi_i = X_j \to a$, when $a \neq \varepsilon$, $g(\pi_i) = a\bar{a}c^2\overline{c^2}\,\bar{c}\,\overline{d^j}\,\bar{c}$; when $a = \varepsilon$, $g(\pi_i) = c^3\overline{c^3}\,\bar{c}\,\overline{d^j}\,\bar{c}$. Then $T = \{g(\pi_1), \cdots, g(\pi_m)\} = g(P)$ is both prefix and suffix set, so $g(P^*) = g(P)^* = T^*$ is a BSR language. Let $w = cdc \in \hat{\Sigma}_2^*$. It can be proved that

$$X_1 \xRightarrow[G\ \text{left}]{\pi_{i_1}\cdots\pi_{i_s}*} uX_{j_1}\cdots X_{j_q},\ u \in \Sigma^* \text{ iff}$$

(i) $wg(\pi_{i_1}\cdots\pi_{i_s}) >^* cd^{j_q}c\cdots cd^{j_1}c$ and

(ii) $h \circ g(\pi_{i_1}\cdots\pi_{i_s}) = u$.

From this fact it follows that $L = h(D_{\Sigma_2} \cap wL_R)$ and h is limited for L, where $L_R = T^*$.

From Theorem 2, we obtain

Corollary 1 AFL $\mathscr{L}_2$ is principal, where $\mathscr{L}_2$ is the class of context-free languages.

3. The deterministic form of Theorem 2

Let $\mathscr{L}_2'$ be the class of deterministic context-free languages, denote

$\mathscr{N} = \{L \mid L$ can be accepted by dpda with empty store$\}$

$\mathscr{T} = \mathscr{L}_2' - \mathscr{N}$.

we know $\mathscr{T} \neq \phi$[7].

Lemma 2[11] Let $L \subseteq \Sigma^*$, L is a deterministic context-free language iff $L \cdot \{\$\} \in \mathscr{N}$, where $\$ \notin \Sigma$.

Now we denote

$\mathscr{N}_0 = \{L \mid L$ can be accepted by a real-time dpda with empty store$\}$

$\mathscr{N}_1 = \{L \mid L \in \mathscr{N} - \mathscr{N}_0$ and $\varepsilon \notin L\}$

Theorem 3 $\mathscr{N}_1 \neq \phi$.

The outline of proof For any $L \subseteq \Sigma^*$, we establish a right-congerence $\sim_L$ on Σ^* as follows: $x \sim_L y$ iff $\forall z \in \Sigma^*$ $(xz \in L \Leftrightarrow yz \in L)$. Then we can prove that if $\mathscr{N}_1 = \phi$, then for any language in $\mathscr{L}_2'$, the relation $\sim_L$ has finite inner-index, it is contrary to that the language $L_0 = \{a^i b^j \mid i \neq j\}$ is in $\mathscr{L}_2'$ and the inner-index of $\sim_{L_0}$ is infinite.

Remark 2 Theorem 3 has its actual interesting. It is well-known, those grammars such as $LR(k)$, $SLR(k)$ etc., their parsing programming can be essentially realized by a dpda with empty store[8],[13],[14]. Naturally, we expect that all of the programming can be realized by a real-time dpda with empty store, Theorem 3 shows that, this is impossible. The necessary condition do that is the inner-index of $\sim_L$ finite.

Theorem 4 (The form of Theorem 2 in "deterministic" case): For any alphabet Σ_1, there exists a alphabet Σ_2, a word $w \in \hat{\Sigma}_2^*$, a Dyck set $D_{\Sigma_2} \subseteq \hat{\Sigma}_2^*$ and a homomorphism h from $\hat{\Sigma}_2^*$ onto Σ_1^* which satisfy the property that for any deterministic context-free L on Σ_1, a regular BSR language $L_R \subseteq \hat{\Sigma}_2^*$ can be found such that $L = h(D_{\Sigma_2} \cap wL_R)$, and h is essential injective for $D_{\Sigma_2} \cap wL_R$.

The outline of proof From Lemma 2 and remark 1, it is easy to construct a grammar G such that for any $w \in L$, there is unique left-most derivation to generate w. By use of the similar method in Theorem 2, we may find the L_R such that $L = h(D_{\Sigma_2} \cap wL_R)$. Then by the uniqueness of the left-most derivation it follows that h is essential injective.

Remark 3 From Theorem 3, we know that h is not necessary to be limited.

The author would like to thank Guo Yu-qi for his critical reading of the manuscript and many valuable suggestions.

References

[1] Book, R. V., Simple representations of certain classes of languages, *J. ACM*, 25: 1 1978.

[2] Ginsburg, S. and Greibach, S., Principal AFL, *J. CSS*, 4, 1970.

[3] Brzozowski, A., Roots of star events, *J. ACM*, 14: 3, 1967.

[4] 郭聿琦，正则语言关于正则P-(S-)语言的一种分解，兰大学报，2，1980.

[5] 郭聿琦，李廉，自由有根语言，左(右)单有根语言以及前者关于后者的自由积分解，数学学报待发表.

[6] Salomaa, A., Formal Languages, Academic Press, New York/ London, 1973.

[7] Hopcroft, J. E. and Ullman, J. D., Formal languages and their relation to automata, Addison—Wesley, Mass. 1969.

[8] Schützenberger, M. P., On context—free languages and pushdown automata, *Inf. and Contr*, 6: 3, 1963.

[9] Knuth, D. E., On the translation of languages from left to right, *Inf. and Contr.* 8: 6, 1965.

[10] Kasai, T., A universal context—free grammar, *Inf. and Contr*, 28: 1, 1975.

[11] Harrison, M. A. and Havel, I. M., Strict deterministic grammars, *J. CSS*, 7: 3, 1973.

[12] Oyamaguchi, M. Inagaki, Y. and Honda, N., A real-time strictness test for deterministic pushdown automata, *Inf. and Contr.*, 47: 1, 1980.

[13] 陈有琪，BCLR (k) 文法及其分析算法，计算机学报. 3: 3, 1980.

[14] Deremer, F., Simple *LR(k)* grammars, *CACM*, 14: 7, 1971.