数理解析研究所講究録
458巻 1982年 309-317

GENERALIZED PARENTHESIS LANGUAGES AND MINIMALIZATION OF THEIR PARENTHESIS PARTS

(extended abstract)

Hideki Yamasaki

Masako Takahashi

Department of Information Science

Tokyo Institute of Technology

1. INTRODUCTION

The parenthesis grammar defined by McNaughton [2] is a context-free grammar $G = (N,K,P,S)$ such that the terminal alphabet K contains a pair of parentheses, say < and >, and the production rules are of form

$$A \rightarrow \langle u \rangle$$

where A is a nonterminal symbol, and u is a word not containing the parentheses < and >. Then for parenthesis grammars the equivalence problem was proved to be decidable [2].

The generalized parenthesis language is defined [3] by extending the spirit of parenthesis languages so that it reflects the block structure prevalent in modern programming languages, while preserving the mathematical wealth.

Let K be an alphabet that includes a set

$$\hat{I} = \{ a, \bar{a} \mid a \text{ is in } I \}$$

of parentheses, and $G = (N,K,P,S)$ be a context-free grammar (cfg, for short) such that the production rules in P are of form

$$A \rightarrow au\bar{a}B, \quad A \rightarrow bB, \text{ or } \quad A \rightarrow e$$

where A and B are in the nonterminal alphabet N, a is in

I, u is a word over $N \cup K$ not containing symbols in I, and b is in $J = K - \hat{I}$. (The e stands for the empty word.) Then we call G a generalized parenthesis grammar (gpg, for short), and the language generated thereby a generalized parenthesis language (gpl, for short) over K with the parenthesis part $\hat{I}$ (or simply, over K[I]).

The class of gpl's so defined has been proved to have nice mathematical features; for example, the equivalence problem for gpg's over K[I] are decidable, and they enjoy various closure properties (under language-theoretic operations in relativized forms, with respect to the 'universal' gpl specified below) [3], [4]. On the other hand, the expressive power of gpl is sufficiently large; for example, it can descrive the syntax of ALGOL 60 with five pairs of parentheses, (,), [,], if, then, begin, end, and ', ' [5].

In this paper, after a short preliminary in the rest of this section, in section 2 we study relations between regular sets and gpl's, and solve some decision problems affirmatively. In particular, we show that the regularity problem for gpl's is decidable, and that for a given regular set L over K and a set $\hat{I}$ of parentheses in K, one can decide whether L is a gpl over K[I] or not. In section 3 we apply these results to the study of parenthesis parts of gpl's, resulting in affirmative answers to more general problems. Among others we prove that for a given gpg G over K[I] and a subset I' of I it is decidable whether L(G) is a gpl over K[I'] or not. Thus we can minimalize the parenthesis part of a given gpl. (If the minimalized parenthesis part is empty then the gpl is regular.) In section 4, relations between gpl's and context-free languages (cfl's, for short) are studied. We give a characterization of cfl's and that of gpl's, both in terms of universal gpl's, regular sets, and projections. We also give a negative answer to the decision problem to ask whether a given cfg generates a gpl or not.

Let $\hat{I} = \{ a, \bar{a} \mid a \text{ is in } I \} \subseteq K$, and $J = K - \hat{I}$ as above. Consider the gpg $G = (\{S\},K,P,S)$ such that

$$P = \{ S \to aS\bar{a}S,\ S \to bS,\ S \to e \mid a \text{ is in } I \text{ and } b \text{ in } J \}.$$

Any gpl over $K[I]$ is included in the gpl generated by G. We call the language $L(G)$ the universal gpl over $K[I]$, and denote it by $D_{I,J}$. In case of $J = \phi$, the language equals the Dyck set D_I over $\hat{I}$. If $I = \phi$ then $D_{I,J} = J^*$. In general, $D_{I,J}$ is equal to $\mathrm{Shuffle}(D_I,J^*)$, the shuffle product of D_I and J^*.

For each element w of $D_{I,J}$, the nonnegative integer $\mathrm{depth}_I(w)$ is defined as follows:

$$\mathrm{depth}_I(e) = 0,$$

$$\mathrm{depth}_I(au\bar{a}v) = \max\{ 1 + \mathrm{depth}_I(u),\ \mathrm{depth}_I(v) \},$$

$$\mathrm{depth}_I(bu) = \mathrm{depth}_I(u).$$

where a is in I, b is in J, and u and v are in $D_{I,J}$. For a language L in $D_{I,J}$, we define

$$\mathrm{depth}_I(L) = \sup\{ \mathrm{depth}_I(w) \mid w \text{ is in } L \},$$

which may or may not be finite.

If uvw is a word in $D_{I,J}$ then we can write

$$v = v_0\bar{a}_1v_1\bar{a}_2v_2\cdots\bar{a}_nv_na_{n+1}v_{n+1}\cdots a_{n+m}v_{n+m}$$

for some $a_1,a_2,\ldots,a_{n+m}$ in I, $v_0,v_1,\ldots,v_{n+m}$ in $D_{I,J}$ and $n,m \geq 0$. In this case we will write

$$|v|_I = \bar{a}_1\bar{a}_2\cdots\bar{a}_na_{n+1}a_{n+2}\cdots a_{n+m}.$$

For a language L in $D_{I,J}$, we define

$\text{subword}_I(L) = \{ v \text{ in } D_{I,J} \mid uvw \text{ is in } L \text{ for some } u,w \}.$

For any word w in $D_{I,J}$, we define the word $\text{surface}_I(w)$ in J^* as follows: If

$$w = u_0(a_1 v_1 \bar{a}_1)u_1(a_2 v_2 \bar{a}_2)u_2 \ldots (a_n v_n \bar{a}_n)u_n$$

for some $n \geq 0$, $u_0, \ldots, u_n$ in J^*, $a_1, \ldots, a_n$ in I, and $v_1, \ldots, v_n$ in $D_{I,J}$, then

$$\text{surface}_I(w) = u_0 u_1 \ldots u_n.$$

For a language L in $D_{I,J}$, we define

$$\text{surface}_I(L) = \{ \text{surface}_I(w) \mid w \text{ is in } L \}.$$

We may suppress the suffix I in these notations when it is clear from the context.

This paper is an extended abstract of [5], and we will omit the proofs of theorems.

2. REGULAR SETS AND GENERALIZED PARENTHESIS LANGUAGES

It has been proved [4] that the class of gpl's over $K[I]$ is closed under intersection with regular sets, and therefore any regular set included in $D_{I,J}$ is a gpl over $K[I]$. In this section we study properties of these regular sets, and give positive answers to some decision problems for gpl's.

Theorem 2.1 If L is a regular set included in $D_{I,J}$, then depth(L) is finite.

Theorem 2.2 If L is a gpl over $K[I]$ and depth(L) is finite, then L is regular.

Corollary 2.3 For a language L in $D_{I,J}$ the following three conditions are equivalent.

(1) L is a regular set.

(2) L is a gpl over K[I], and depth(L) is finite.

(3) L is obtained from subsets of J by a finite number of applications of regular operations $\cup$, $\cdot$, $*$, and bracketting by symbols in I (i.e., $aX\bar{a}$ for X, where a is in I).

Theorem 2.4 For a given regular expression E over K and a set of parentheses $\hat{I}$ in K, one can decide whether the regular set L denoted by E is a gpl over K[I]. If this is the case, one can effectively obtain a gpg over K[I] to generate the set L.

Note that any regular set in K* is a gpl over $K[\phi]$. Therefore to specify the parenthesis part $\hat{I}$ in theorem 2.4 is important. From the theorem, for a given regular set L in K*, we can effectively list up all the paired subalphabets $\hat{I}$ of K such that L is a gpl with parenthesis part $\hat{I}$.

Theorem 2.5 Whether a given gpg generates a regular set or not is decidable.

3. ON MINIMALIZATION OF THE PARENTHESIS PART

The regularity problem for gpg's (theorem 2.5) is nothing but to ask whether the parenthesis part of a given gpg can be reduced to the empty set. In this section we consider a more general problem to minimalize the parenthesis part of a given gpg. First we note a property of the mapping surface $:D_{I,J} \to J^*$.

Theorem 3.1 If L is a gpl over $K[I]$, then surface(L) is a regular set over $K-\hat{I}$.

As a consequence we know that if a gpl L over $K[I]$ is also a gpl over $K[I']$ where $I' \subseteq I$ then $\text{surface}_{I'}(L)$ is a regular subset of $D_{I-I',J}$. The converse of this statement is not true. However we can prove the following.

Theorem 3.2 Let $G = (N,K,P,S)$ be a gpg over $K[I]$, $L_A = \{ w$ in $K^* \mid A \overset{*}{\Rightarrow} w$ in $G \}$ for each A in N, and $I' \subseteq I$. If $\text{surface}_{I'}(L_A)$ is regular for each A, then each L_A is a gpl over $K[I']$.

Theorem 3.3 Let L be a gpl over $K[I]$, and $I' \subseteq I$. Then L is a gpl over $K[I']$ if and only if $\text{surface}_{I'}(\text{subword}_I(L))$ is regular (i.e., $\text{depth}_{I-I'}(\text{surface}_{I'}(\text{subword}_I(L)))$ is finite).

Corollary 3.4 Let G be a gpg over $K[I]$, and $I' \subseteq I$. Then it is decidable whether $L(G)$ is a gpl over $K[I']$ or not. If the answer is affirmative, one can effectively obtain a gpg G' over $K[I']$ to generate the language $L(G)$.

As for the expansion of the parenthesis parts of gpl's, we can extend theorem 2.4 as follows.

Theorem 3.5 Let L be a gpl over $K[I]$, $I \subseteq I'$, and $\hat{I}' \subseteq K$.

Then L is a gpl over $K[I']$ if and only if $L \subseteq D_{I',J'}$ where $J' = K - \hat{I}'$. For a given gpg G over $K[I]$ and an expansion I' of I, one can decide whether the condition is satisfied or not. If this is the case one can effectively obtain a gpg G' over $K[I']$ to generate the language $L(G)$.

By corollary 3.4 and theorem 3.5, for a given gpg G over $K[I]$ we can list up all restrictions and expansions I' of I such that $L(G)$ is a gpl over $K[I']$. In particular, we can obtain all minimal parenthesis parts for a given gpg G over $K[I]$, i.e., all minimal subsets I' of I such that $L(G)$ is a gpl over $K[I']$.

It is interesting to note that a gpl may have no 'minimum' nor 'maximum' parenthesis part. For instance, consider

$$L = \{ (ab)^i(cd)^i \mid i=0,1,2,\ldots \}.$$

The language L is a gpl in various ways; it is a gpl with $\hat{I}_1 \sim \hat{I}_5$ below as the parenthesis part.

$$\hat{I}_1 = \{ a, d \},$$
$$\hat{I}_2 = \{ a, c \},$$
$$\hat{I}_3 = \{ b, c \},$$
$$\hat{I}_4 = \{ b, d \},$$
$$\hat{I}_5 = \{ a, d; b, c \}.$$

Among these, $\hat{I}_1 \sim \hat{I}_4$ are minimal, while $\hat{I}_2$, $\hat{I}_4$ and $\hat{I}_5$ are maximal. But none of them is the minimum, nor the maximum. At present we do not know any algorithm to get all possible parenthesis parts of a given gpl, or whether one can expect it at all or not.

4. CONTEXT-FREE LANGUAGES AND GENERALIZED PARENTHESIS LANGUAGES

First we note that any gpl is a deterministic cfl, indeed Greibach and Friedman [1] have shown a stronger result that any gpl is a superdeterministic language.

We prove that a language L is a cfl (or gpl, respectively) if and only if $L = h(D_{I,J} \cap R)$ for a regular set R and a projection (or pair preserving projection) h. Finally we prove the undecidability of whether a given cfl is a gpl or not.

Let K and K' be alphabets. A homomorphism $h : K^* \to K'^*$ is said to be a projection if $h(K) \subseteq K'$. A projection $h : (\hat{I} \cup J)^* \to (\hat{I}' \cup J')^*$ is said to be pair preserving if $h(I) \subseteq I'$, $h(J) \subseteq J'$ and $h(\bar{a}) = \overline{h(a)}$ for any a in I.

Theorem 4.1 A language L is a gpl over K[I] if and only if $L = h(D_{I',J'} \cap R)$ for some alphabets I' and J', a pair preserving projection h, and a regular set R over $\hat{I}' \cup J'$.

Theorem 4.2 A language L is a cfl if and only if $L = h(L')$ for a gpl L' over some alphabet K[I], and a projection h.

Corollary 4.3 A language L is a cfl if and only if $L = h(D_{I,J} \cap R)$ for a projection h and a regular set R over some alphabet $K = \hat{I} \cup J$.

Theorem 4.4 For a given cfg whether it generates a gpl or not is undecidable.

REFERENCES

[1] Greibach,S.A., and Friedman,E.P., Superdeterministic PDAs: A subcase with a decidable inclusion problem, JACM Vol.27, 675-700 (1980).

[2] McNaughton,R., Parenthesis grammars, JACM Vol.14, 490-500 (1967).

[3] Takahashi,M., Generalizations of regular sets and their application to a study of context-free languages, Inf. & Control, Vol.27, 1-35 (1975).

[4] Takahashi,M., Nest sets and relativized closure properties, Research Reports on Information Science, C-35 (1981), Tokyo Institute of Technology.

[5] Yamasaki,H. and Takahashi,M., Generalized parenthesis languages and minimalization of their parenthesis parts, Research Reports on Information Science, C-39 (1982), Tokyo Institute of Technology.