CANONICAL SCATTERED CONTEXT GENERATORS OF SENTENCES WITH THEIR PARSES

Jiří TECHET, Master Degree Programme (5)
Dept. of Information Systems, FIT, BUT
E-mail: xteche00@stud.fit.vutbr.cz

Supervised by: Dr. Alexander Meduna

ABSTRACT

The scattered context generators derive their sentences followed by the corresponding parses. This paper discusses their canonical version, which makes this derivation in a leftmost way. It demonstrates that for every recursively enumerable language, L, there exists a canonical scattered context generator whose language consists of L's sentences followed by their parses. In fact, this result is established based on the generators containing no more than six nonterminals.

1 INTRODUCTION

Very recently, the scattered context grammars without erasing productions have been used to generate their sentences together with the corresponding parses in [1]. Recall that it was demonstrated that for every recursively enumerable language, L, there exists a scattered context grammar whose language consists of L's sentences followed by their parses. In this paper, we define the *proper leftmost generator of its sentences with their parses* which makes its generation by making only the leftmost derivation, and we demonstrate the characterization of recursively enumerable languages by analogy with the characterization described above. Moreover, this grammar contains at most six nonterminals.

2 PRELIMINARIES

We assume that the reader is familiar with the language theory (see [2]). V^* represents the free monoid generated by V under the operation of concatenation. The unit of V^* is denoted by ε. Set $V^+ = V^* - \{\varepsilon\}$. For $w \in V^*$, $|w|$ and $alph(w)$ denote the length of w and the set of symbols occurring in w, respectively. For $L \subseteq V^*$, $alph(L) = \{a \,|\, a \in alph(w), w \in L\}$.

A *queue grammar* is a sixtuple, $Q = (V, T, W, F, s, P)$, where V and W are alphabets satisfying $s \in VW$, $T \subseteq V$, $F \subseteq W$, $s \in (V-T)(W-F)$, and $P \subseteq (V \times (W-F)) \times (V^* \times W)$

is a finite relation whose elements are called productions. For every $a \in V$, there exists a production $(a,b,x,c) \in P$. If $u,v \in V^*W$ such that $u = arb$, $v = rxc$, $a \in V$, $r,x \in V^*$, $b,c \in W$, and $(a,b,x,c) \in P$, then $u \Rightarrow v\ [(a,b,x,c)]$ in G or, simply, $u \Rightarrow v$. In the standard manner, extend $\Rightarrow$ to $\Rightarrow^n$, where $n \geq 0$; then, based on $\Rightarrow^n$, define $\Rightarrow^+$ and $\Rightarrow^*$. The language of Q, $L(Q)$, is defined as $L(Q) = \{w \in T^* \mid s \Rightarrow^* wf \text{ where } f \in F\}$. It is proved that for every queue grammar, Q', there exists an equivalent queue grammar, Q, $L(Q') = L(Q)$, such that its generation can be divided into two parts: first, it generates only words over $(V - T)$, and, after it passes through a special state, it generates only words over T.

3 DEFINITIONS

A *scattered context grammar*, a *SCG* for short, is a quadruple, $G = (V,P,S,T)$, where V is an alphabet, $T \subseteq V$, $S \in V - T$, and P is a finite set of productions such that each production has the form $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)$, for some $n \geq 1$, where $A_i \in V - T$, $x_i \in V^*$, for $1 \leq i \leq n$. If every production $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n) \in P$ satisfies $x_i \in V^+$ for all $1 \leq i \leq n$, G is a *propagating scattered context grammar*, a *PSCG* for short. If $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n) \in P$, $u = u_1A_1u_2\ldots u_nA_nu_{n+1}$, and $v = u_1x_1u_2\ldots u_nx_nu_{n+1}$, where $u_i \in V^*$, $1 \leq i \leq n$, then G makes a *derivation step* from u to v according to $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)$, symbolically written as $u \Rightarrow v\ [(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)]$ in G or, simply, $u \Rightarrow v$; in addition, if $A_i \notin alph(u_i)$ for all $1 \leq i \leq n$, then this step is *leftmost*. The *language of* G is denoted by $L(G)$ and defined as $L(G) = \{x \mid x \in T^*, S \Rightarrow^* x\}$. If $S \Rightarrow^* x$ with $x \in T^*$, $S \Rightarrow^* x$ is a successful generation of x in G. If every step in every successful generation in G is leftmost, G *generates* $L(G)$ *in a leftmost way*.

In this paper, we automatically assume that for every grammar, G, there is a *set of production labels*, $lab(G)$, such that its cardinality is equal to the number of G's productions and no member of $lab(G)$ occurs in any of G's components. Furthermore, there is a bijection from the set of G's productions to $lab(G)$ such that if this bijection maps a production $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)$ to a label $l \in lab(G)$, we say that $(A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)$ *is labeled by* l, symbolically written as $l : (A_1,\ldots,A_n) \rightarrow (x_1,\ldots,x_n)$. By analogy with labeling each production in every *SCG*, we label each production, (a,b,x,c), in every queue grammar as $l : (a,b,x,c)$. To express that G makes $x \Rightarrow^* y$ by using a sequence of productions labeled with $p_1,p_2,\ldots,p_n$, we write $x \Rightarrow^* y\ [\rho]$, where $x,y \in V^*$, $\rho = p_1\ldots p_n \in lab(G)^*$. Let $S \Rightarrow^* x\ [\rho]$ in G, where $x \in T^*$ and $\rho \in lab(G)^*$; then, x is a *sentence generated by* G *according to parse* ρ. Let $G = (V,P,S,T)$ be a *SCG* with $lab(G) \subseteq T$. G is a *proper generator of its sentences with their parses* if $L(G) = \{x \mid x = y\rho, y \in (T - lab(G))^*, \rho \in lab(G)^*, S \Rightarrow^* x\ [\rho]\}$; in addition, if G generates $L(G)$ in a leftmost way, G is a *proper leftmost generator of its sentences with their parses*.

4 RESULTS

To define the main theorem formally, we introduce the weak identity π from $(V \cup lab(G))^*$ to V^* defined as $\pi(a) = a$ for every $a \in V$ and $\pi(p) = \varepsilon$ for every $p \in lab(G)$.

Theorem 1. *For every recursively enumerable language, L, there exists a PSCG, G, such that G contains no more than six nonterminals, G is a proper leftmost generator of its*

sentences with their parses and $L = \pi(L(G))$.

Proof. Let $Q = (V,T,W,F,s,R)$ be a queue grammar such that $L = L(Q)$. Define an injection, α, from $lab(Q)$ to $\{\bar{0}\}^*\{\bar{1}\}$ so that α is an injective homomorphism when its domain is extended to $lab(Q)^*$ in the standard way. Similarly, define an injection, β, from T to $\{0\}^*\{1\}$, so that β is an injective homomorphism when its domain is extended to T^*. Further, define the binary relation, f, over V so that $f(\varepsilon) = \varepsilon$ and $f(a) = \{\alpha(r) \mid r : (a,b,c_1 \dots c_n,d) \in R\}$ for all $a \in V$. Similarly, define the binary relation, g, over W so that $g(b) = \{\alpha(r) \mid r : (a,b,c_1 \dots c_n,d) \in R\}$ for all $b \in W$. In the standard manner, extend the domain of f and g to V^* and W^*, respectively. Define the *PSCG* as $G = (\{S,A,B,\#,\bar{0},\bar{1}\},P,S,\{0,1\} \cup lab(G))$, where P and $lab(G)$ are constructed as follows:

1. For every $\bar{a}_0 \in f(a_0)$, $\bar{q}_0 \in g(q_0)$ such that $s = a_0q_0$, add
   $\lfloor 1\bar{a}_0\bar{q}_0 \rfloor : (S) \to (A \lfloor 1\bar{a}_0\bar{q}_0 \rfloor AA\bar{q}_0A\bar{a}_0AB)$ to P;

2. For every $r : (a,b,c_1 \dots c_n,d) \in R$, $c_1,\dots,c_n \in (V-T)$ for some $n \geq 0$ and $d \in (W-F)$, $\bar{c}_1 \in f(c_1) \dots \bar{c}_n \in f(c_n)$, $\bar{d} \in g(d)$, add
   $\lfloor 2r\bar{c}_1 \dots \bar{c}_n\bar{d} \rfloor : (A,A,A,A,A,B) \to (A, \lfloor 2r\bar{c}_1 \dots \bar{c}_n\bar{d} \rfloor A, \alpha(r)A, \bar{d}A, \bar{c}_1 \dots \bar{c}_nA, B)$ to P;

3. Add $\lfloor 3 \rfloor : (A,A,A,A,A,B) \to (A, \lfloor 3 \rfloor A, A, A, B, A)$ to P;

4. For every $r : (a,b,c_1 \dots c_n,d) \in R$, $c_1,\dots,c_n \in T$ for some $n \geq 0$ and $d \in (W-F)$, $\bar{d} \in g(d)$, add
   $\lfloor 4r\bar{d} \rfloor : (A,A,A,A,B,A) \to (\beta(c_1) \dots \beta(c_n)A, \lfloor 4r\bar{d} \rfloor A, \alpha(r)A, \bar{d}A, B, A)$ to P;

5. For every $r : (a,b,c_1 \dots c_n,d) \in R$, $c_1,\dots,c_n \in T$ for some $n \geq 0$ and $d \in F$, add
   $\lfloor 5r \rfloor : (A,A,A,A,B,A) \to (\beta(c_1) \dots \beta(c_n), \lfloor 5r \rfloor A, \alpha(r)A, A, B, AA)$ to P;

6. Add
   $\lfloor 6 \rfloor : (A,\bar{0},A,\bar{0},A,\bar{0},B,A,A) \to (\lfloor 6 \rfloor, A, \#, A, \#, A, B, A, A)$, and
   $\lfloor 7 \rfloor : (A,\bar{1},A,\bar{1},A,\bar{1},B,A,A) \to (\lfloor 7 \rfloor, A, \#, A, \#, A, B, A, A)$ to P;

7. Add
   $\lfloor 8 \rfloor : (A,A,A,B,A,A) \to (\lfloor 8 \rfloor B, \#, \#, \#, \#, \#)$,
   $\lfloor 9 \rfloor : (B,\#) \to (\lfloor 9 \rfloor, B)$, and
   $\lfloor 10 \rfloor : (B) \to (\lfloor 10 \rfloor)$ to P.

Every sentence $w \in L(G)$ is generated in G in this way: $S \Rightarrow x_1\ [\lfloor 1\bar{a}_0\bar{q}_0 \rfloor] \Rightarrow^* x_2\ [\rho] \Rightarrow y\ [\lfloor 3 \rfloor] \Rightarrow^* z\ [\sigma] \Rightarrow u\ [\lfloor 5r \rfloor] \Rightarrow^* v\ [\tau] \Rightarrow w_1\ [\lfloor 8 \rfloor] \Rightarrow^* w_2\ [\omega] \Rightarrow w\ [\lfloor 10 \rfloor]$ where ρ, σ and τ, ω are sequences consisting of labels of the productions introduced in steps 2, 4 of the construction, and $\{\lfloor 6 \rfloor, \lfloor 7 \rfloor\}$, $\{\lfloor 9 \rfloor\}$, respectively.

REFERENCES

[1] Meduna, A., Techet, J.: Generation of Sentences with Their Parses: the Case of Propagating Scattered Context Grammars, Acta Cybernetica 17 (2005), (in press).

[2] Rozenberg, G., Salomaa, A.: Handbook of Formal Languages, Volume 1 through 3, Springer-Verlag, 1997.