Unambiguous Equal Matrix Languages

RANI SIROMONEY

Department of Mathematics, Madras Christian College, Madras 59, India

A necessary and sufficient algebraic condition is given for a subclass of equal matrix languages to be unambiguous. It is shown that all languages from this class are unambiguous. Recursive unsolvability of ambiguity in equal matrix grammar and decidability of ambiguity in bounded equal matrix grammar are proved.

1. INTRODUCTION

In view of the importance of unambiguity in programming languages, questions of ambiguity and inherent ambiguity have been extensively studied (Ginsburg and Ullian, 1966). For those context-free languages (CFL) which are known to be inherently ambiguous, it is possible to write unambiguous equal matrix grammars and further equal matrix languages (EML) have the property that the corresponding Parikh mapping is semilinear (Siromoney, 1969a, 1969b). EML also correspond to k-tuples of words accepted by k-tape one-way nonwriting automaton extensively studied (Rabin and Scott, 1959; Elgot and Mezei, 1964; Rosenberg, 1967; Fischer and Rosenberg, 1968).

In this paper, we establish a necessary and sufficient algebraic condition for EML contained in $a_1^* \ldots a_k^*$ to be unambiguous and prove that all languages from this class are unambiguous. This is interesting in view of the existence of inherently ambiguous CFL in $a_1^*a_2^*a_3^*$ (Chomsky and Schutzenberger, 1963), $a_1^*a_2^*a_3^*a_4^*$ (Parikh, 1966) and $a_1^*a_2^*a_3^*a_4^*a_5^*$ (Ginsburg, 1966), and the fact that for subsets of $a_1^* \ldots a_k^*$, the generative power of EMG is higher than that of CFG (Siromoney, 1969a).

In Section 2, we obtain a necessary and sufficient algebraic condition for EML subsets of $a_1^* \ldots a_k^*$ to be unambiguous and establish that every EML in $a_1^* \ldots a_k^*$ is unambiguous. Section 3 is concerned with establishing that it is recursively unsolvable whether an arbitrary EMG is ambiguous and showing that, in contrast to the full family of EML, it is decidable whether an arbitrary EMG for a bounded set is ambiguous. These two results are similar to those for context-free grammars.
The terminology is basically that of Ginsburg (1966) and all undefined terms and notation are as in Ginsburg and Ullian (1966).

**Definition 1.1.** An equal matrix grammar (EMG) of order \( k \) is a 4-tuple \( G = (V, I, P, S) \) where

1. \( V \) consists of the alphabet \( I \), a set \( V_N \) of nonterminals consisting of the initial symbol \( S \), and a set of disjoint \( k \)-tuples \( \langle A_1, \ldots, A_k \rangle \).
2. \( P \) consists of the following types of matrix rules:
   1. a set of initial matrix rules (initial rules) of the form
      \[
      [S \rightarrow f_1 A_1 \cdots f_k A_k],
      \]
      where \( f_1, \ldots, f_k \) are in \( I^* \), \( S \) the initial symbol and \( \langle A_1, \ldots, A_k \rangle \) in \( V_N \).
   2. a set of nonterminal equal matrix rules (nonterminal rules) of the form
      \[
      \begin{bmatrix}
      A_1 & f_1 B_1 \\
      \vdots & \vdots \\
      A_k & f_k B_k
      \end{bmatrix},
      \]
      where \( f_1, \ldots, f_k \) are in \( I^* \) and \( \langle A_1, \ldots, A_k \rangle, \langle B_1, \ldots, B_k \rangle \) in \( V_N \).
   3. a set of terminal equal matrix rules (terminal rules) of the form
      \[
      \begin{bmatrix}
      A_1 & f_1 \\
      \vdots & \vdots \\
      A_k & f_k
      \end{bmatrix}
      \]
      where \( f_1, \ldots, f_k \) are in \( I^* \) and \( \langle A_1, \ldots, A_k \rangle \) in \( V_N \).

An equal matrix grammar (EMG) is an EMG of any finite order \( k \).

**Definition 1.2.** \( L \subseteq I^* \) is an equal matrix language if there is an EMG \( G = (V, I, P, S) \) such that \( L = L(G) \) where \( L(G) = \{ w \in I^* | S \Rightarrow^* w \} \). \( L(G) \) is said to be the language generated by \( G \).

**Definition 1.3.** Given an EMG, \( G = (V, I, P, S) \), a word \( w \) in \( L(G) \) is said to be ambiguously derivable if there exist more than one derivation of \( w \) from \( S \). It can easily be seen from the definition of EMG and the nature of the derivation of a word using the rules of an EMG, that this will be the case if and only if the associated generation trees are different. An EMG \( G \) is said
to be ambiguous if there exists a word in \(L(G)\) derived ambiguously from \(S\), and is otherwise unambiguous. An EML \(L\) is said to be unambiguous if there is at least one unambiguous EMG generating \(L\).

A set \(L \subseteq I^*\) is said to be bounded, if there is a \(k\) such that every word \(w\) in \(L\) is of the form \(w_1^{(1)} \cdots w_k^{(k)}\) and an EMG generating a bounded set is called a bounded EMG.

2. Unambiguity of EML in \(a_1^* \cdots a_k^*\)

In this section, we give a necessary and sufficient algebraic condition for EML in \(a_1^* \cdots a_k^*\) to be unambiguous and thus establish that every EML in \(a_1^* \cdots a_k^*\) has at least one unambiguous EMG generating it.

**Theorem 2.1.** Let \(I = \{a_j/1 \leq j \leq k\}\) and \(a = \langle a_1, \ldots, a_k \rangle\). If \(M\) is an EML contained in \(a_1^* \cdots a_k^*\) and \(f_a^{-1}(M)\) is the finite union of disjoint linear sets \(L_i\) each with independent periods (FUDLIP) then \(M\) has an unambiguous grammar.

**Proof.** \(M\) is an EML contained in \(a_1^* \cdots a_k^*\) if and only if \(L = f_a^{-1}(M)\) is semilinear (Siromoney, 1969a). Since \(f_a\) is one to one, \(M\) is the finite union of disjoint languages \(f_a(L_i)\) and it is enough to show that each \(f_a(L_i)\) has an unambiguous grammar. We explicitly construct an unambiguous EMG \(G\) to generate \(M\) where \(f_a^{-1}(M)\) is a linear set \(L = L(c; p_1, \ldots, p_n)\). Let \(c = (c_1, \ldots, c_k), p_i = (p_{i1}, \ldots, p_{in}), i = 1, \ldots, n\). Then \(G = (V, I, P, S)\) where \(I = \{a_i/i = 1, \ldots, k\}\), \(V - I - S = \langle A_{i1}, \ldots, A_{ki} \rangle/i = 1, \ldots, n\) and \(P\) consists of the rules:

\[
\begin{align*}
[S & \rightarrow A_1 \cdots A_k] & [S & \rightarrow a_i^{e_i}A_{i1} \cdots a_k^{e_k}A_{ki}] & i = 1, \ldots, n \\
\{A_1 & \rightarrow a_1^{e_1}\}, & \{A_{i1} & \rightarrow a_i^{p_{i1}}A_{i1}\}, & \{A_{kt} & \rightarrow a_k^{p_{kt}}\} & i = 1, \ldots, n \\
A_k & \rightarrow a_k^{e_k}, & A_{ki} & \rightarrow a_k^{p_{ki}}A_{ki} & i = 1, \ldots, n - 1, \quad j = 2, \ldots, n & j > i \\
A_{ki} & \rightarrow a_k^{p_{ki}}A_{kj} & & & \\
\end{align*}
\]

It is easily seen that \(G\) is an unambiguous EMG generating \(L\).

**Theorem 2.2.** Let \(I = \{a_j/1 \leq j \leq k\}\) and \(a = \langle a_1, \ldots, a_k \rangle\). If a set \(M\) contained in \(a_1^* \cdots a_k^*\) has an unambiguous EMG then \(f_a^{-2}(M)\) is a FUDLIP.
Proof. The proof parallels that of Lemma 5.2 of Ginsburg and Ullian (1966) and we sketch the main steps omitting details.

Let $G = (V, I, P, S)$ be an unambiguous EMG generating $M$ where $M$ is infinite, $G$ is $e$-free and "reduced".

1. A period of $\langle A \rangle = \langle A_1, ..., A_k \rangle$ in $V - I - S$ is defined to be $(u_1, ..., u_k)$ if $A_1 \ldots A_k \Rightarrow u_1 A_1 \ldots u_k A_k$ where $u_1, ..., u_k$ are in $I^*$ and $u_1 \ldots u_k \neq e$. It can be seen that $u_i$ is in $a_i^*$, $i = 1, ..., k$ (from the proof of Lemma 2.2 of Siromoney, 1969b).

2. The minimal period of $\langle A \rangle$ is defined and it can be shown that there can be at most one minimal period for each $k$-tuple $\langle A \rangle$, (otherwise the hypothesis of unambiguity will be contradicted) and that if $(a_1^{(1)}, ..., a_k^{(k)})$ be the minimal period of $\langle A \rangle$ with the unique derivation

$$A_1 \ldots A_k \Rightarrow a_1^{(1)} A_1 \ldots a_k^{(1)} A_k \Rightarrow \ldots \Rightarrow a_1^{(n)} A_1 \ldots a_k^{(n)} A_k$$

then each of $\langle A_{1i}, ..., A_{ki} \rangle$, $i = 1, ..., n$ has $(a_1^{(1)}, ..., a_k^{(k)})$ as its minimal period.

3. Define $U_{\langle A \rangle} = \{\langle A \rangle, \langle A \rangle | i = 1, ..., n\}$, i.e., $U_{\langle A \rangle}$ is such that every $k$-tuple in $U_{\langle A \rangle}$ has the same minimal period. It is enough to consider $M'$ generated by $V'$ a subset of $V$ such that all elements in $V'$ are used in the generation of elements of $M'$. $\tau$ is taken to be the extension of $f_a^{-1}$ to include nonterminals and $(a_1^{(1)}, ..., a_k^{(k)})$, $i = 1, ..., m$ the minimal periods of $k$-tuples $\langle A_i \rangle$ with $U_{\langle A \rangle} \subseteq V'$. Let $\nu_i = \tau(a_1^{(1)}, ..., a_k^{(k)})$, $i = 1, ..., m$. Let $J$ be the set of all words in $I^*$ generated from $S$ by using only variables from $V'$ and that too only once or twice and let $z_1, ..., z_s$ be those words $z$ in $J$ such that $\tau(z) \neq \tau(w) + \nu_i$, $1 \leq i \leq m, w$ in $J$. Let

$$u_i = \tau(z_i) \quad \text{and} \quad Q_i = \{u_i + \sum_{j=1}^m k_j \nu_j | k_j \geq 0\}, \quad 1 \leq i \leq s.$$

4. To complete the proof it is enough to show that (a) $\tau(M') = \bigcup_{i=1}^s Q_i$, (b) each $Q_i$ has linearly independent set of periods and (c) the $Q_i$'s are pairwise disjoint.

(a). The proof is a modification of Parikh's argument for semi-linearity of CFL and this method has been extended to EML in Siromoney (1969a) by replacing single variables by $k$-tuples of variables. In the same
manner the proof of (a) in Ginsburg and Ullian (1966) can be modified by replacing variables by $k$-tuples of variables.

(b) If the periods of $Q_i$ are not linearly independent then one of the $v$'s say $v_1$ is a linear combination of the others, i.e., $v_1 = k_2v_2 + \cdots + k_rv_r$. Due to the minimality of the periods of each $k$-tuple involved, this will imply that there are two distinct derivations of the same word, contradicting the fact that $G$ is unambiguous.

(c) Finally to show that the $Q_i$'s are pairwise disjoint, it is clear again that if two $Q_i$'s say $Q_i$ and $Q_j$ are not disjoint the elements in $Q_i \cap Q_j$ will give rise to ambiguously derivable words contradicting the original hypothesis that $G$ is unambiguous.

Combining Theorems 2.1 and 2.2 we have the following result. Let $I = \{a_j/1 \leq j \leq k\}$ and $a = \langle a_1, \ldots, a_k \rangle$. An EML $M$ contained in $a_1^* \cdots a_k^*$ is unambiguous iff $f^{-1}_a(M)$ is a FUDLIP.

**Corollary 2.1.** Every EML in $a_1^* \cdots a_k^*$ is unambiguous.

Proof follows from the fact that every semilinear set is a FUDLIP (Eilenberg and Schützenberger, 1969) and from Theorems 2.1 and 2.2.

### 3. Decision Problems

In this section we first establish the recursive unsolvability of the ambiguity of an arbitrary EMG following the methods of Landweber (1964) and then show that the same problem for an arbitrary bounded EMG does have a decision procedure.

Landweber (1964) has reduced the classical Post's correspondence problem for a fixed alphabet $I$ and integer $n$ to a problem concerning $Q(x) = \{b_{i(1)} \cdots b_{i(r)}x_{i(1)} \cdots x_{i(r)}/1 \leq i(j) \leq n, 1 \leq j \leq r, r \geq 1\}$ where $I_0 = \{b_1, \ldots, b_n\}$ are new distinct symbols, $I_1 = I \cup I_0$ and $(x) = (x_1, \ldots, x_n)$ is a list of $n$ $I$-words. Thus if $(x) = (x_1, \ldots, x_n)$ and $(y) = (y_1, \ldots, y_n)$ are lists of $n$ $I$-words then the correspondence problem for these lists has a solution if and only if $Q(x) \cap Q(y)$ is empty. $Q(x)$ is a CFL over $I_1$. We define $E(x) = \{b_{i(1)} \cdots b_{i(r)}x_{i(1)} \cdots x_{i(r)}/1 \leq i(j) \leq n, 1 \leq j \leq r, r \geq 1\}$ over $I_1$ and it can be shown that there is an unambiguous EMG generating $E(x)$. If $(x) = (x_1, \ldots, x_n)$ and $(y) = (y_1, \ldots, y_n)$ are lists of $n$ $I$-words, then the
correspondence problem for these lists has a solution iff \( E(x) \cap E(y) \) is empty. 

\( G \) consisting of the rules

\[
[S \rightarrow A_1B_1], \quad [A_1 \rightarrow b_iA_1] \quad [A_1 \rightarrow b_i] \quad i = 1, \ldots, n
\]

\[
[B_1 \rightarrow x_iB_1], \quad [B_1 \rightarrow x_i] \quad i = 1, \ldots, n
\]

\[
[S \rightarrow A_2B_2], \quad [A_2 \rightarrow b_iA_2] \quad [A_2 \rightarrow b_i] \quad i = 1, \ldots, n
\]

\[
[B_2 \rightarrow y_iB_2], \quad [B_2 \rightarrow y_i] \quad i = 1, \ldots, n
\]

generates \( E(x) \cup E(y) \) unambiguously iff \( E(x) \cap E(y) \) is empty. This leads to

**Theorem 3.1.** It is recursively unsolvable to decide whether an arbitrary EMG \( G \) is unambiguous.

Next we show that the same problem for an arbitrary bounded EMG does have a decision procedure.

The following lemma can be obtained by modifying the proof of Lemma 3.5 of Ginsburg and Ullian (1966).

**Lemma 3.1.** Let \( L \) be a bounded EML, each word of which contains exactly one occurrence of the ordered \( h \)-tuple \( c_1, \ldots, c_h \) (i.e. every word in \( L \) is of the form \( u_1c_1 \cdots u_d c_h \) where \( u_1, \ldots, u_d \) do not contain any of the letters \( c_1, \ldots, c_h \)). If \( M \subseteq w_1^* \cdots w_k^* \) is an EML then \( \{u_1w_1^{(1)} \cdots u_kw_k^{(h)} | u_1c_1 \cdots u_d c_h \text{ in } L \} \) is a bounded EML.

**Theorem 3.2.** Given a bounded EMG \( G \), the set of words ambiguously derivable in \( G \) is a bounded EML and effectively calculable.

**Proof.** Let \( G = (V, I, P, S) \) be an EMG of order \( k \) for a bounded set \( L \subseteq w_1^* \cdots w_k^* \). We can test to see if \( L \neq \emptyset \). If \( L = \emptyset \) the theorem is trivial. Suppose \( L \neq \emptyset \). We may assume that (1) each \( k \)-tuple in \( G \) generates at least one word in \( I^* \) and (2) for each \( k \)-tuple \( \langle A_1, \ldots, A_k \rangle \) in \( V - I - S \), \( S \) generates a word \( x_1A_1 \cdots x_kA_k \), \( x_1, \ldots, x_k \) in \( I^* \). Let \( S \) and \( \langle A_1, \ldots, A_k \rangle, \ldots, \langle A_{1n}, \ldots, A_{kn} \rangle \) be the distinct variables of \( G \). For \( 1 \leq i \leq n \), let \( w_{i,1}, \ldots, w_{i,m(i)} \) be all the distinct words \( w \) in \( V^* \) such that \( A_{1i} \cdots A_{ki} \Rightarrow w \) and let \( w_1, \ldots, w_m \) be all the words \( w \) in \( V^* \) such that \( S \Rightarrow w \). (We note that the \( w \)'s are of the form \( x_1A_1 \cdots x_kA_k \), \( x_1, \ldots, x_k \) in \( I^* \) and \( A_1, \ldots, A_k \) in \( V - I - S \)). Let \( L(\psi) = \{x | x \not\Rightarrow \psi, \text{ } x \in I^* \} \) where \( \psi = S \) or \( x_{1i}A_{1i} \cdots x_{ki}A_{ki} \), \( i = 1, \ldots, n \). Let \( L(w_i), \ldots, L(w_{i,m(i)}) \) be all bounded EML. Hence \( \Gamma = \bigcup_{j \neq k} (L(w_j) \cap L(w_k)) \) is a bounded EML and

\[
\Gamma_i = \bigcup_{j \neq k} (L(w_{i,j}) \cap L(w_{i,k}))
\]
is a bounded EML for each $i$. For $1 \leq i \leq n$, let $L_i$ be the set of all words $x$ of the form $x_{1i}A_{1i} \cdots x_{ki}A_{ki}$ such that $S \Rightarrow x$. Each $L_i$ is a bounded EML. Let $Q_i$ be the set of all words $u_{1i}w_{1i} \cdots u_{ki}w_{ki}$ with $w_{1i} \cdots w_{ki}$ in $\Gamma_i$, $u_{1i} \cdots u_{ki}$ in $I^*$ such that $u_{1i}A_{1i} \cdots u_{ki}A_{ki}$ is in $L_i$. By Lemma 4.1 each $Q_i$ is semilinear. Hence $\bigcup_{i=1}^{n} Q_i$ is a bounded EML. It is shown that $\Gamma \cup (\bigcup_{i=1}^{n} Q_i)$ is the set of words ambiguously derivable in $G$. The theorem will then follow since $\Gamma \cup (\bigcup_{i=1}^{n} Q_i)$ is effectively calculable.

Suppose $x$ is in $\Gamma$. Then $x$ is in some $L(w_i)$ and $L(w_j)$, $w_i \neq w_j$. Hence there exist two distinct derivations $S \Rightarrow w_i \Rightarrow^* x$ and $S \Rightarrow w_j \Rightarrow^* x$. Therefore $x$ is ambiguously derivable in $G$. Suppose $x$ is in $Q_i$ for some $i$. Then there exist $w = x_{1i} \cdots x_{ki}$ in $\Gamma_i$, $u_1, \ldots, u_k$ in $I$ such that $u_{1i}A_{1i} \cdots u_{ki}A_{ki}$ is in $L_i$ and $x = u_{1i}x_{1i} \cdots u_{ki}x_{ki}$. Since $u_{1i}A_{1i} \cdots u_{ki}A_{ki}$ is in $L_i$, there exists at least one derivation $S \Rightarrow u_{1i}A_{1i} \cdots u_{ki}A_{ki}$. Since $w$ is in $\Gamma_i$ there are distinct $j$ and $l$ such that $x_{1i} \cdots x_{ki}$ and $u_{1i}A_{1i} \cdots u_{ki}A_{ki}$ are the productions used at level $p$ in $T_1$ and $T_2$, respectively. Clearly $w$ is in $L(w_i) \cap L(w_j)$ and hence in $\Gamma$. (b) $p > 2$. The generation trees up to level $(p - 1)$ are identical in $T_1$ and $T_2$. At level $(p - 1)$ let $\langle A_{1i}, \ldots, A_{ki} \rangle$ be the $k$-tuple involved. Then there exist disjoint $j$ and $l$ such that $A_{1i} \cdots A_{ki} \Rightarrow w_{jl}$ and $A_{1i} \cdots A_{ki} \Rightarrow w_{jl}$ are the productions used at level $p$ in $T_1$ and $T_2$, respectively. Also there exist subtrees $S_1$ of $T_1$ and $S_2$ of $T_2$ such that $A_{1i} \cdots A_{ki}$ generates $w_1 \cdots w_k$ in $I^*$. Clearly $w_1 \cdots w_k$ is in $L(w_{1i}) \cap L(w_{1j})$ and hence in $\Gamma_i$. Hence $x$ is in $Q_i$. Thus $\Gamma \cup (\bigcup_{i=1}^{n} Q_i)$ is the set of ambiguously derivable words in $G$.

**Corollary.** It is solvable whether an arbitrary bounded EMG is ambiguous.
References


ROSENBERG, A. L. (1967), A machine realization of the linear context-free languages, Information and Control 10, 175–188.
