Title: An Automaton for Deciding Whether a Given Set of Words is a Code. (Algebraic Semigroups, Formal Languages and Computation)

Author(s): Tsuji, Kayoko

Citation: 数理解析研究所講究録 (2001), 1222: 123-127

Issue Date: 2001-07

URL: http://hdl.handle.net/2433/41325

Type: Departmental Bulletin Paper

Textversion: publisher

Kyoto University
An Automaton for Deciding Whether a Given Set of Words is a Code.

天理大学教養部 江村代子 (Kayoko Tsuji)
Faculty of Liberal Arts,
Tenri University

For a given finite set \( X \) of words on an alphabet \( A \), it is well-known that there is an algorithm for deciding whether the set \( X \) is a code (see [1]). In this paper, we define the ambiguous word which has more than two factorizations in \( X^* \) and we construct an automaton \( \mathcal{A}_X \) such that the set \( L(\mathcal{A}_X) \) recognized by \( \mathcal{A}_X \) is the set of all ambiguous words in \( \mathcal{A}_X \). We show that a given set \( X \) of words on an alphabet \( A \) is a code if and only if that the set \( L(\mathcal{A}_X) \) recognized by \( \mathcal{A}_X \) is empty.

For a finite set \( X \) of words on an alphabet \( A \), we denote by \( P(X) \) the set \( \{ p \in X \mid p \) is a proper prefix of some word in \( X \} \). Let \( c_1 \) be the cardinality of \( P(X) \). Then, there is an injection \( \varphi \) from \( P(X) \) into the set of natural numbers such that \( 1 \leq \varphi(p) \leq c_1 \) for every \( p \in P(X) \). We also denote by \( S(X) \) the set \( \{ s \in A^* \mid s \) is a proper suffix of some word in \( X \} \). There is an injection \( \psi \) from \( S(X) \) into the set of natural numbers such that \( c_1 + 1 \leq \psi(s) \) for every \( s \in S(X) \).

Now, we construct an automaton \( \mathcal{A}_X \) over \( A \) inductively. The edges and states of \( \mathcal{A}_X \) are defined by the following rules. Let \( i \) be the unique initial state of \( \mathcal{A}_X \). Every element of \( \varphi(P(X)) \) is state of \( \mathcal{A}_X \) and, for every word \( p \) in \( P(X) \),

\[
\varphi(p) \xrightarrow{\varphi(p)}
\]

is a path in \( \mathcal{A}_X \). As every word \( p \) in \( P(X) \) is a proper prefix of some word \( x \) in \( X \), the word \( u = p^{-1}x \) is a suffix of \( x \), that is, \( u \) is in \( S(X) \). Then, \( \psi(u) \) is a state of \( \mathcal{A}_X \) and

\[
\varphi(p) \xrightarrow{u} \psi(u)
\]

is a path in \( \mathcal{A}_X \). If \( \psi(u) \) is a state of \( \mathcal{A}_X \) and if, for some word \( v \) in \( S(X) \), the concatenation
$uv$ is written of the form
\[ uv = x_1 \cdots x_m \] where $x_1, \ldots, x_m$ is in $A^*$ and $v$ is a proper suffix of $x_m$ then, $\psi(v)$ is a state of $A^*$ and
\[ \psi(u) \rightarrow \psi(v) \]
is a path in $A^*$. Since $\varphi(P(X)) \cup \psi(S(X))$ and $S(X)$ are finite, this procedure has finite steps. Let $Q$ be the set of all states of $A^*$. The set of terminal states of $A^*$ is the set $\psi(S(X) \cap X^*) \cap Q$.

A word $w$ is said to be ambiguous if there exist words $x_1, \ldots, x_m, y_1, \ldots, y_n$ of $X$ such that
\[ w = x_1 \cdots x_m = y_1 \cdots y_n \] and $x_1 \cdots x_i \neq y_1 \cdots y_j$ $(i = 1, \ldots, m - 1, j = 1, \ldots, n - 1)$.

**Theorem.** For a given set $X$ on an alphabet $A$, the set $L(A^*)$ recognized by the automaton $A^*$ of $X$ is the set of all ambiguous words in $X^*$.

**Proof.** Let $w \in L(A^*)$. There exist $x_1 \in P(X), x_2, \ldots, x_r \in S(X), x_r \in S(X) \cap X^*$ such that $w = x_1x_2 \cdots x_r$ and a succesible path
\[ i \xrightarrow{x_1} q_1 \xrightarrow{x_2} q_2 \xrightarrow{x_3} \cdots q_{r-1} \xrightarrow{x_r} q_r \]
where $q_1 = \varphi(x_1), q_2 = \psi(x_2), \ldots, q_r = \psi(x_r)$ and $q_r$ is a terminal state. Since $q_1 = \varphi(x_1)$ is in $\varphi(P(X))$, $q_1$ is not a terminal state. If $r = 2$, then
\[ i \xrightarrow{x_1} q_1 \xrightarrow{x_2} q_2 \]
is succesible. By the definition of $A^*$ the word $w = x_1x_2$ itself is in $X$. Thus, $w$ is ambiguous. Let $r > 2$. By the definition of $A^*$, $x_1, x_2, \ldots, x_r$ $(k = 2, \ldots, r)$ are words of $X^*$, thus
\[ w = x_1x_2 \cdots x_r \]
has two factorizations:
\[ w = y_1y_2 \cdots y_m = z_1z_2 \cdots z_n \] $(y_1, \ldots, y_m, z_1, \ldots, z_n \in X)$.

We show that $y_iy_2 \cdots y_j = z_1z_2 \cdots z_j$ for all $i = 1, \ldots, m, j = 1, \ldots, n$. Suppose that
\[ y_iy_2 \cdots y_j = z_1z_2 \cdots z_j \] for some $i, j$. There exists $x_i$ such that $y_iy_2 \cdots y_i = z_1z_2 \cdots z_j$ is a prefix of $x_i x_2 \cdots x_r$, and that $y_iy_2 \cdots y_i = z_1z_2 \cdots z_j$ is not a prefix of $x_i x_2 \cdots x_{r-i}$. By the definition of $A^*$, we may assume that $y_i$ is a subwords of $x_{r-i}$, then $x_i$ is a suffix of $y_i$. However, $z_j$ must be a subwords of $x_i x_{r+i}$. It is impossible.
Let \( w \) be ambiguous and \( w = y_1 y_2 \cdots y_m = z_1 z_2 \cdots z_n \) \((y_1, \ldots, y_m, z_1, \ldots, z_n \in X)\), \( y_i y_2 \cdots y_{j-1} = z_i z_2 \cdots z_j \) for all \( i = 1, \ldots, m \), \( j = 1, \ldots, n \). We may assume that \( z_1 \) is a proper prefix of \( y_1 \). Since \( w \) is ambiguous, there exist \( i_1, i_2, \ldots, j_1, j_2, \ldots \) such that the following conditions are satisfied:

- \( y_1 \) is a proper prefix of \( z_2 \cdots z_{j_1} \) and not a prefix of \( z_2 \cdots z_{j_1-1} \)
- \( z_2 \cdots z_{j_1} \) is a proper prefix of \( y_1 \cdots y_{i_1} \) and not a prefix of \( y_1 \cdots y_{i_1-1} \)

Let \( x_1 = x_1^{-1} y_1, x_2 = x_1^{-1} z_2 \cdots z_{j_1}, x_3 = (z_2 \cdots z_{j_1})^{-1} y_2 y_3 \cdots y_n \), \( \ldots \). If \( z_n \) is a proper prefix of \( y_m \) and if \( z_2 z_3 \cdots z_m \) is a suffix of \( y_m \) and \( z_2 z_3 \cdots z_m \) is not, then we set

\[ x_r = (z_2 \cdots z_{j_1})^{-1} y_2 \cdots y_m = z_k \cdots z_n \].

If \( y_m \) is a proper prefix of \( z_n \) and if \( y_k y_{k+1} \cdots y_m \) is a suffix of \( z_n \) and \( y_k y_{k+1} \cdots y_m \) is not, then we set \( x_r = (y_2 \cdots y_{k-1})^{-1} z_2 \cdots z_n = y_k \cdots y_m \).

Then, we have a susccesible path

\[ i \to x_1 \to q_1 \to x_2 \to q_2 \to \cdots \to x_{r-1} \to q_r \]

where \( q_1 = \varphi(x_1), q_2 = \psi(x_2), \ldots, q_r = \psi(x_r) \). q.e.d.

The proposition yields the following corollary immediately.

**Corollary.** A given set \( X \) on an alphabet \( A \) is a code if and only if the set \( L(\mathcal{A}_X) \) recognized by the automaton \( \mathcal{A}_X \) of \( X \) is empty.

**Example 1.** Let \( A = \{a, b\} \) and let \( X = \{aab, aabb, aabbab, aba, baa, bbaaba\} \). We construct \( \mathcal{A}_X \) and we show that \( L(\mathcal{A}_X) \) is empty, therefore, \( X \) is a code.

It is clear that \( P(X) = \{aab, aabb\} \). We define a bijection \( \varphi : P(X) \to \{1, 2\} \) by \( \varphi(aab) = 1, \varphi(aabb) = 2 \). Since \( aab \) is a prefix of \( aabb \), \( aabbab \) and \( aabb \) is a prefix of \( aabbab \), then \( b, bab, ab \) are in \( S(X) \). We can define an injection \( \psi \) from \( S(X) \) into the set of natural numbers such that \( \psi(b) = 3, \psi(bab) = 4, \psi(ab) = 5 \). Since \( b \) is a prefix of \( baa \), \( bbaaba \) and \( ab \) is a prefix of \( aba \), then \( aa, baaba, a \) are in \( S(X) \). But, \( bab \) is not prefix of any word of \( X \), therefore the state \( \psi(bab) = 4 \) is not coaccessible. Continuing this process, we have the following automaton:
where $\psi(aa) = 6$, $\psi(b aba) = 7$, $\psi(bb b) = 8$, $\psi(a) = 9$, $\psi(ba) = 10$, $\psi(abb) = 11$, $\psi(bb) = 12$, $\psi(a aba) = 13$, $\psi(abb a b) = 14$. Since $S(X) \cap X = \{aba\}$ and $\psi(b a b)$ is not a state of $A_X$, there is no terminal state in $A_X$, thus $L(A_X)$ is empty and $X$ is a code.

**Example 2.** Let $A = \{a, b\}$ and let $X = \{aaaaba, aba, a bab, bab, babaa, bba\}$. We construct $A_X$. In this case $L(A_X)$ is not empty, therefore, $X$ is not a code. It is clear that $P(X) = \{aba, bab\}$. We define a bijection $\varphi : P(X) \rightarrow \{1, 2\}$ by $\varphi(aba) = 1, \varphi(bab) = 2$. We have the following automaton:

The set of all states $Q$ is $\{\psi(aba) = 1, \psi(bab) = 2, \psi(b) = 3, \psi(abaaa) = 4, \psi(ab) = 5, \psi(ba) = 6, \psi(aaa) = 7, \psi(aba) = 8, \psi(a) = 9, \psi(b a a a) = 10, \psi(aaba) = 11, \psi(b a b) = 12\}$. The set of all
terminal states is $S(X) \cap X \cap Q = \{ \psi(aba) = 8, \psi(bab) = 12 \}$. Since $\psi(aba) = 8$ is a terminal state, a path

\[
\begin{align*}
1 & \xrightarrow{aba} 2 \xrightarrow{b} 3 \xrightarrow{ba} 6 \xrightarrow{b} 3 \xrightarrow{ab} 5 \xrightarrow{ab} 5 \xrightarrow{a} 9 \xrightarrow{bab} 12 \xrightarrow{aaa} 7 \xrightarrow{aba} 8
\end{align*}
\]

is successful, thus $w = ababbababababaaaaba$ is accepted by $A_X$ and $w$ is ambiguous. In fact, $w$ has two factorizations

\[
w = (aba)(bba)(bab)(aba)(babaa)(aba) = (abab)(bab)(abab)(aba)(aaaaba)
\]

in $X$.

Reference