LETTER

A Note on Pcodes of Partial Words

Tetsuo MORIYA[†a)] *and* **Itaru KATAOKA**[†], ***Members***

SUMMARY In this paper, we study partial words in relation with pcodes, compatibility, and containment. First, we introduce $C_{\subset}(L)$, the set of all partial words contained by elements of L, and $C_{\supset}(L)$, the set of all partial words containing elements of L, for a set L of partial words. We discuss the relation between $C(L)$, the set of all partial words compatible with elements of the set L, $C_{\subset}(L)$, and $C_{\supset}(L)$. Next, we consider the condition for $C(L)$, $C_{\subset}(L)$, and $C_{\supset}(L)$ to be a pcode when L is a pcode. Furthermore, we introduce some classes of pcodes. An infix pcode and a comma-free pcode are defined, and the inclusion relation among these classes is established.

key words: *formal language, partial word, pcode, compatible*

1. Introduction

Partial words are strings over a finite alphabet that may contain a number of "do not know" symbols. The motivation behind the notion of partial words is the comparison of two genes (or two proteins). Alignment of two such strings can be viewed as a construction of two partial words that are said to be compatible in a sense that will be described in Sect. 2.

Codes play an important role in the study of combinatorics on words [1], [9]. In [4], pcodes were introduced in relation with combinatorics on partial words. While a code L of words does not allow two distinct decipherings of some word in L^+, a pcode K of partial words does not allow two distinct "compatible" decipherings in K^+.

Some combinatorial properties of partial words have been investigated in previous studies [2]–[5], [7], [8], [10].

In this paper, we study partial words in relation with pcodes, compatibility, and containment. Let L be a set of partial words. In [6], the set $C(L)$ of all partial words compatible with the elements of a set L of partial words was defined.

We introduce the following two sets of partial words in relation with $C(L)$.

(1) $C_{\subset}(L)$, the set of all partial words containing elements of L, and

(2) $C_{\supset}(L)$, the set of all partial words contained by elements of L.

First, we discuss the relation between $C(L)$, $C_{\subset}(L)$, and $C_{\supset}(L)$. Next, we consider the condition for $C(L)$, $C_{\subset}(L)$, and $C_{\supset}(L)$ to be a pcode when L is a pcode. Furthermore, we introduce some classes of pcodes. An infix pcode and a comma-free pcode are defined, and the inclusion relation among these classes is established.

2. Preliminaries

Let Σ be a nonempty finite set of symbols, which we call an alphabet. A word over the alphabet Σ is a finite sequence of elements of Σ. The empty sequence is called an *empty word* and is denoted by ϵ. The set of all words over Σ is denoted by Σ^*. The set of nonempty words over Σ is denoted by Σ^+. Thus, $\Sigma^+ = \Sigma^* \backslash \{\epsilon\}$.

For w in Σ^*, $|w|$ denotes the length of w. A *language* over Σ is a set $L \subseteq \Sigma^*$.

A word of length n over Σ can be defined by a total function $u : \{0, 1, \ldots, n-1\} \to \Sigma$ and is usually represented as $u = a_0 a_1 \ldots a_{n-1}$ with $a_i \in \Sigma$.

A partial word of length n over Σ is a partial function $u : \{0, 1, \ldots, n-1\} \to \Sigma$. For $0 \leq i < n$, if u is defined, then we say that i belongs to the domain of u (denoted by $i \in D(u)$); otherwise, we say that i belongs to the set of holes of u (denoted by $i \in H(u)$). A word over Σ is a partial word over Σ with an empty set of holes (we refer to words as *full words*). For any partial word u over Σ, $|u|$ denotes its length. Clearly, $|\epsilon| = 0$. Let $W_0(\Sigma)$ denote the set Σ^*, and for $i \geq 1$, let $W_i(\Sigma)$ denote the set of partial words over Σ with at most i holes. We put $W(\Sigma) = \cup_{i \geq 1} W_i(\Sigma)$, the set of all partial words over Σ with an arbitrary number of holes.

If u is a partial word of length n over Σ, then the companion of u (denoted by $u_\diamond$) is the total function $u_\diamond : \{0, 1, \ldots, n-1\} \to \Sigma \cup \{\diamond\}$ defined as

$u_\diamond = u(i)$ if $i \in D(u)$, $\diamond$ otherwise.

The symbol $\diamond \notin \Sigma$ is considered the "do not know" symbol. The word $u = ab\diamond ab\diamond a$ is the companion of the partial word u of length 7, where $D(u) = \{0, 1, 3, 4, 6\}$ and $H(u) = \{2, 5\}$. The bijectivity of the map $u \mapsto u_\diamond$ allows us to define partial words concepts such as concatenation and powers, in a trivial manner. The set $W(\Sigma)$ is a monoid under the concatenation of partial words (ϵ serves an identity). For convenience in the sequel, we say, for instance, "the partial word $ab\diamond ab\diamond a$" instead of "the partial word with companion $ab\diamond ab\diamond a$".

Given two subsets L, K of $W(\Sigma)$, we define $LK = \{uv | u \in L \text{ and } v \in K\}$. We sometimes write $L \sqsubset K$ if $L \subset K$ but $L \neq K$.

A factorization of a partial word u is any sequence $u_1, u_2, \ldots, u_i$ of partial words such that $u = u_1 u_2 \ldots u_i$. For a subset L of $W(A)$ and integer $i \geq 0$, let L^i denote the set $\{u_1 u_2 \ldots u_i | u_1, \ldots, u_i \in L\}$. For a subset L of $W(\Sigma)$, we use

Manuscript received July 26, 2013.

Manuscript revised September 19, 2013.

[†]The authors are with School of Science and Engineering, Kokushikan University, Tokyo, 154–8515 Japan.

a) E-mail: moriya@kokushikan.ac.jp

DOI: 10.1587/transinf.E97.D.139

the notation $\|L\|$ for the cardinality of L.

Let L^* denote the submonoid of $W(\Sigma)$ generated by L, or $L^* = \bigcup_{i\geq 0} L^i$, where $L^0 = \{\epsilon\}$, and let L^+ denote the subsemigroup of $W(\Sigma)$ generated by L, or $L^+ = \bigcup_{i>0} L^i$. An element of $\{\diamond\}^+$ is called a *holeword*. If u and v are partial words of equal length, then u is said to be *contained* in v, denoted by $u \subset v$ or $v \supset u$ if all elements in $D(u)$ are in $D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. We sometimes write $u \sqsubset v$ if $u \subset v$ but $u \neq v$. The partial words u and v are compatible, denoted by $u \uparrow v$ if there exists a partial word w such that $u \subset w$ and $v \subset w$. Let $u \vee v$ denote the least upper bound of u and v.

Let $L \subseteq W(\Sigma)$. We define $C(L)$, $C_\subset(L)$, and $C_\supset(L)$ as follows:

$C(L) = \{y \in W(\Sigma)|\ x \uparrow y$ for some $x \in L\}$.
$C_\subset(L) = \{y \in W(\Sigma)|\ x \subset y$ for some x $\in L\}$.
$C_\supset(L) = \{y \in W(\Sigma)|\ x \supset y$ for some x $\in L\}$.

Let L be a nonempty subset of $W(\Sigma)\backslash\{\epsilon\}$. Then, L is a *pcode* if for all integers $m \geq 1, n \geq 1$ and partial words $u_1, \ldots, u_m, v_1, \ldots, v_n \in L$, the condition

$u_1 \ldots u_m \uparrow v_1 \ldots v_n$

implies that $m = n$ and $u_i = v_i$ for $i = 1, \ldots, m$.

3. $C(L)$, $C_\subset(L)$, and $C_\supset(L)$

In this section, first, we discuss the relation between $C(L)$, $C_\subset(L)$ and $C_\supset(L)$ for a set L of partial words.

Proposition 1: For $L \subseteq W(\Sigma)$, $C(L) = C_\supset(C_\subset(L))$.

Proof. Let $y \in C(L)$. There exists $x \in L$ such that $x \uparrow y$, that is, there exists $z \in W(\Sigma)$ such that $x \subset z$ and $y \subset z$. It follows that $z \in C_\subset(L)$ and that $y \in C_\supset(z) \subseteq C_\supset(C_\subset(L))$. Thus, $C(L) \subseteq C_\supset(C_\subset(L))$.

Conversely, let $z \in C_\supset(C_\subset(L))$. There exist $x \in L$ and $y \in W(\Sigma)$ such that $x \subset y$ and $z \subset y$. We have $x \uparrow z$. Hence, $z \in C(L)$. Thus, $C_\supset(C_\subset(L)) \subseteq C(L)$. ::

Next, we consider the condition for $C(L)$, $C_\subset(L))$, and $C_\supset(L)$ to be a pcode when L is a pcode.

Proposition 2: Let $L \subseteq W(\Sigma)\backslash\epsilon$.
1. $C_\subset(L)$ is a pcode iff $L \subseteq \Sigma^*$ and L is a pcode.
2. $C_\supset(L)$ is a pcode iff L is equal to a singleton set of a holeword.
3. $C(L)$ is a pcode iff $L \subseteq \Sigma^*$ and L is a pcode.

Proof.
1.[If] If $L \subseteq \Sigma^*$, then $C_\subset(L)$=L. Thus, the result holds.
[Only if] Suppose that $L \not\subseteq \Sigma^*$. Then, there exists $x \in L$ such that $\|H(x)\| \geq 1$. Moreover, there exists $y \in W(\Sigma)$ such that $y \in C_\subset(L)$, $x \sqsubset y$, and $x \uparrow y$. Since $x \in C_\subset(L)$, it follows that $C_\subset(L)$ is not a pcode. Next, suppose that L is not a pcode. Since $L \subseteq C_\subset(L)$, $C_\subset(L)$ is not a pcode.
2.[If] Trivial.
[Only if] Suppose that L is not a singleton set of holewords. (Case 1) L is a set of holewords. ($L \subseteq \{\diamond\}^+$.) Then there exist two distinct holewords x and y. Then, we have $xy \uparrow yx$. Since $x, y \in C_\supset(L)$, it follows that $C_\supset(L)$ is not a pcode.
(Case 2) L is not a set of holewords. There exists $x \in W(\Sigma)\backslash\{\diamond\}^+$ such that $x \in L$. Then $x \uparrow y$ where $y = (\diamond)^{|x|}$. Since $x, y \in C_\supset(L)$, it follows that $C_\supset(L)$ is not a pcode.
3. The result can be proved as in 1. ::

4. Prefix Pcodes, Infix Pcodes, Comma-Free Pcodes

In this section, we introduce some classes of pcodes.
Let L be a subset of $W(\Sigma)$. The set L is a *prefix pcode* if for all $u, v \in L$, $ux \uparrow v$ for some $x \in W(\Sigma)$ implies that $u = v$ [6]. *Suffix pcodes* are defined in a symmetric manner. *Bifix pcodes* are pcodes that are both prefix and suffix.

The set L is an *infix pcode* if for all $x, y \in W(\Sigma)$ and $u, v \in L$, $v \uparrow xuy$ implies that $u = v$.

The set L is a *comma-free pcode* if for all $u, v, w \in L$ and $x, y \in W(\Sigma)$, $uv \uparrow xwy$ implies that $x = \epsilon$ or $y = \epsilon$.

Proposition 3: Every infix pcode is a bifix pcode.

Proof. Let L be an infix pcode. Suppose that L is not a prefix pcode. Then, for some $u, v \in L$, and $x \in W(\Sigma)\backslash\{\epsilon\}$, $ux \uparrow v$. Since $x \neq \epsilon$, it follows that $u \neq v$. Hence, L is not an infix pcode. This is a contradiction. Thus, L is a prefix pcode. Similarly, we can prove that L is a suffix pcode. Hence, L is a bifix pcode. ::

Remark 1: A bifix pcode is not necessarily an infix pcode. For example, consider a bifix pcode $L_1 = \{a \diamond a, b\}$. Note that $a \diamond a \uparrow aba$.

Proposition 4: Every comma-free pcode is an infix pcode.

Proof. Let $L \subseteq W(\Sigma)$ be a comma-free pcode. Assume that L is not an infix pcode. Then, there exist $u, v \in L$, $x, y \in W(\Sigma)$, such that $v \uparrow xuy$ and $xy \neq \epsilon$. This implies that $vv \uparrow xuyxuy$. Since $xy \neq \epsilon$, it follows that $xuyx \neq \epsilon$ and $y \neq \epsilon$, or $x \neq \epsilon$ and $yxuy \neq \epsilon$. Then, L is not a comma-free pcode. This is a contradiction. Hence, L is an infix pcode. ::

Remark 2: An infix pcode is not necessarily a comma-free pcode. For example, consider an infix pcode $L_2 = \{ab \diamond b, ba \diamond a\}$. Note that $ab \diamond bab \diamond b \uparrow aba \diamond ab \diamond b$.

References

[1] J. Berstel and D. Perrin, Theory of Codes, Academic Press, New York, 1985.

[2] J. Berstel and L. Boasson, "Partial words and a theorem of Fine and Wilf," Theor. Comput. Sci., vol.218, pp.135–141, 1999.

[3] F. Blanchet-Sadri, "Periodicity on partial words," Computers and Mathematics with Applications, vol.47, pp.71–82, 2004.

[4] F. Blanchet-Sadri, "Codes, orderings, and partial words," Theor. Comput. Sci., vol.329, pp.177–202, 2004.

[5] F. Blanchet-Sadri and Ajay Chriscoe, "Local periods and binary partial words: an algorithm," Theor. Comput. Sci., vol.314, pp.189–216, 2004. http://www.uncg.edu/mat/AlgBin/

[6] F. Blanchet-Sadri and M. Moorefield. "Pcodes of partial words," Preprint (www.uncg.edu/mat/pcode).

[7] F. Blanchet-Sadri and S. Duncan, "Partial words and the critical factorization theorem," J. Combinatorial Theory, Series A 109, pp.221–245, 2005. http://www.uncg.edu/mat/cft/

[8] F. Blanchet-Sadri and Robert A. Hegstrom, "Partial words and a

theorem of Fine and Wilf revisited," Theor. Comput. Sci., vol.270, pp.401–419, 2002.

[9] H.J. Shyr, Free monoids and languages, Hon Min Book Company, Taichung, Taiwan, 2001

[10] F. Blanchet-Sadri and D.K. Luhmann, "Conjugacy on partial words," Theor. Comput. Sci., vol.289, pp.297–312, 2002.