Journal of Discrete Algorithms 14 (2012) 55–64

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Number of holes in unavoidable sets of partial words I ☆

F. Blanchet-Sadri [a,*], Bob Chen [b], Aleksandar Chakarov [c]

[a] Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170, USA
[b] Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, Dept 0112, LaJolla, CA 92093-0112, USA
[c] Department of Computer Science, University of Colorado at Boulder, 430 UCB, Boulder, CO 80309-0430, USA

ARTICLE INFO

Article history:
Available online 7 December 2011

Keywords:
Automata and formal languages
Combinatorics on words
Partial words
Unavoidable sets

ABSTRACT

Partial words are sequences over a finite alphabet that may contain some undefined positions called holes. We consider unavoidable sets of partial words of equal length. We compute the minimum number of holes in sets of size three over a binary alphabet (summed over all partial words in the sets). We also construct all sets that achieve this minimum. This is a step towards the difficult problem of fully characterizing all unavoidable sets of partial words of size three.

1. Introduction

An *unavoidable* set of (full) words X over a finite alphabet A is one for which every two-sided infinite word over A has a factor in X (when a word w has no factor in X, we say that w avoids X). For example, the set $X = \{aa, ba, bb\}$ is unavoidable over the alphabet $\{a, b\}$, since avoiding aa and bb forces a word to be an alternating sequence of a's and b's. This fundamental concept was explicitly introduced in 1983 in connection with an attempt to characterize the rational languages among the context-free ones [8]. Since then it has been consistently studied by researchers in both mathematics and theoretical computer science (see for example [5–7,9,10,12–14]).

Partial words are sequences that may contain some undefined positions called holes, denoted by $\diamond$'s, that match every letter of the alphabet (we also say that $\diamond$ is *compatible* with each letter of the alphabet). For instance, $a\diamond bca\diamond b$ is a partial word with two holes over $\{a, b, c\}$, while $aabcabb$ is a full word over $\{a, b, c\}$ built by filling in the first hole with an a and the second hole with a b. A set of partial words X over A is unavoidable if every two-sided infinite full word over A has a factor compatible with an element in X.

Unavoidable sets of partial words were introduced in [2], where the problem of characterizing such sets of cardinality n over a k-letter alphabet was initiated. Note that if X is unavoidable, then every two-sided infinite unary word has a factor compatible with a member of X; thus X cannot have fewer elements than the alphabet, and so $k \leqslant n$ (note that the cases $n = 1$ and $k = 1$ are trivial). The characterization of *all* unavoidable sets of cardinality $n = 2$ was settled recently in [3] using deep arguments related to Cayley graphs. So our next long-term goal is to characterize unavoidable sets of cardinality $n = 3$. Since in [2], all such sets over a three-letter alphabet were completely characterized (in fact, there are no non-trivial such sets), we need to focus on sets over a two-letter alphabet.

☆ This material is based upon work supported by the National Science Foundation under Grant No. DMS-0754154. The Department of Defense is also gratefully acknowledged. Part of this paper was presented at IWOCA 2010 (Blanchet-Sadri et al., 2011) [4]. A research assignment from the University of North Carolina at Greensboro for the first author is gratefully acknowledged. Some of this assignment was spent at the LIAFA: Laboratoire d'Informatique Algorithmique: Fondements et Applications of Université Paris 7 – Denis Diderot, Paris, France.

* Corresponding author.
E-mail address: blanchet@uncg.edu (F. Blanchet-Sadri).

1570-8667/$ – see front matter
doi:10.1016/j.jda.2011.12.001

In [2], a complete characterization of all three-word unavoidable sets over a binary alphabet where each partial word has at most two defined positions was given, and some special cases where one partial word has more than two defined positions were discussed, but general criteria for these sets had not been found. In this paper, among other things, we answer affirmatively a conjecture that was left open there. Our main goal however is to make another step towards the full $n = 3$ characterization by computing the minimum number of holes in any unavoidable set of partial words of equal length and of cardinality three over a binary alphabet. We also construct all sets that achieve this minimum.

Our paper is organized as follows: In Section 2, we present the basic definitions and terminology regarding unavoidable sets of partial words that we use throughout the paper. In Section 3, we formally state our main goal towards the major problem on unavoidable sets we are concerned with, that is, the *characterization problem* or the problem of characterizing unavoidable sets of partial words of cardinality n over a k-letter alphabet. In Sections 6 and 7, we make two steps towards this problem. More specifically, our first step is that we give an answer to the above mentioned conjecture on unavoidable sets of size three, while our second step is that we also compute the minimum number of holes in unavoidable sets of size three based on our characterization of these sets in two families (given in Sections 4 and 5). Finally in Section 8, we conclude with some remarks.

2. Unavoidable sets of partial words

In this section, we present the basics on unavoidable sets of partial words together with the notation that we use throughout the paper. We refer the reader to Ref. [1] for more background material.

Let A be a fixed non-empty finite set called an *alphabet* whose elements we refer to as *letters*. A *finite* (*full*) *word* w over A is a finite sequence of letters of A. The sequence of length zero, or the *empty word*, is denoted by ε. We write $|w|$ to denote the length of w, and $w(i)$ to denote the letter at position i. By convention, we begin indexing the positions with 0, so a word w of length m can be represented as $w = w(0)\cdots w(m-1)$. Formally, a finite word of length m is a function $w:\{0,\ldots,m-1\} \to A$. The number of occurrences of the letter a in w is denoted by $|w|_a$. We denote by A^* the set of all finite words over A.

A *two-sided infinite* (*full*) *word* w over A is a function $w:\mathbb{Z} \to A$. For a positive integer p, w is *p-periodic* or is of *period* p, if $w(i) = w(i+p)$ for all $i \in \mathbb{Z}$. We say w is *periodic* if it has a period. If v is a non-empty finite word, then $v^{\mathbb{Z}}$ denotes the unique two-sided infinite word w with period $|v|$ such that $v = w(0)\cdots w(|v|-1)$. Similarly, a *one-sided infinite* (*full*) *word* w over A is a function $w:\mathbb{N} \to A$. A finite word u is a *factor* of w if some integer i satisfies $u = w(i)\cdots w(i+|u|-1)$. An *m-factor* is a factor of length m.

A *partial word* w of length m over A is a function $w:\{0,\ldots,m-1\} \to A_\diamond$, where $A_\diamond = A \cup \{\diamond\}$ with $\diamond \notin A$. For $0 \leqslant i < |w|$, if $w(i) \in A$, then i belongs to the *domain* of w, denoted by $D(w)$. Otherwise, i is in the *set of holes* of w, denoted by $H(w)$. We denote by $A_\diamond^*$ the set of all words over $A_\diamond$ (i.e. the set of all partial words over A, including the empty word, ε). Note that full words are simply partial words without holes, that is, partial words whose domain is the entire set $\{0,\ldots,|w|-1\}$. Two partial words u and v of equal length are *compatible*, denoted by $u \uparrow v$, if $u(i) = v(i)$ whenever $i \in D(u) \cap D(v)$. In this sense, we may view a hole as a "wildcard" character that can match any letter in A. We denote by $h(w)$ the number of holes in w, thus, $h(w) = |w| - |D(w)|$.

Let w be a two-sided infinite word and let u be a partial word. We say w *meets* u if w has a factor compatible with u, and w *avoids* u otherwise. Now, w meets a set of partial words X if it meets some $u \in X$, and w avoids X otherwise. If X is avoided by some two-sided infinite word, then X is *avoidable*; otherwise, X is *unavoidable* or every two-sided infinite word has a factor compatible with an element in X. For example, the set $X = \{a, b\diamond b\}$ is unavoidable over $\{a, b\}$, since avoiding a forces a word to be a sequence of b's. We say X is *m-uniform* if every partial word in X has length m.

The partial word u is *contained* in the partial word v, denoted by $u \subset v$, if $|u| = |v|$ and $u(i) = v(i)$, for all $i \in D(u)$. We say that v is a *strengthening* of u if v has a factor containing u, and write $v \succ u$ (in other words, v has a factor built by "filling in" a number of holes in u). We also say that u is a *weakening* of v. The following illustrates an example:

$$\begin{aligned} u &= b\ \diamond\ \diamond\ \diamond\ a, \\ v &= b\ a\ b\ \diamond\ \diamond\ a\ a\ b\ b\ b. \end{aligned}$$

Note that if a two-sided infinite word w meets the partial word v, it also meets every weakening of v, and if w avoids u then w avoids every strengthening of u.

Let X, Y be sets of partial words. We extend the notions of strengthening and weakening as follows. We say that X is a strengthening of Y (written as $X \succ Y$) if, for each $v \in X$, there exists $u \in Y$ such that $v \succ u$. We also say that Y is a weakening of X. For example,

$$X = \{b\diamond aab, bab\diamond\diamond aabbb\} \succ Y = \{b\diamond\diamond\diamond a, b\diamond a\diamond b, aa\}.$$

It is not hard to see that if the two-sided infinite word w meets X, then it also meets every weakening of X, and if w avoids X then it avoids any strengthening of X. Hence if X is unavoidable, so are all weakenings of X, and if X is avoidable all strengthenings of X are avoidable.

Two partial words u and v are *conjugate*, denoted by $u \sim v$, if there exist partial words x, y such that $u \subset xy$ and $v \subset yx$. It is well known that conjugacy on full words is an equivalence relation, and we use $c(m, k)$ to denote the number of