Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Universal partial words over non-binary alphabets

Bennet Goeckner [a,1], Corbin Groothuis [b,*,3], Cyrus Hettle [c,2], Brian Kell [d,1], Pamela Kirkpatrick [e,2], Rachel Kirsch [b,3], Ryan Solava [f,2]

[a] Department of Mathematics, University of Kansas, Lawrence, KS 66045-7594, USA
[b] Department of Mathematics, University of Nebraska, Lincoln, NE 68588-0130, USA
[c] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA
[d] Google, 6425 Penn Avenue, Pittsburgh, PA 15206, USA
[e] Department of Mathematics, Lehigh University, Bethlehem, PA 18015, USA
[f] Department of Mathematics, Vanderbilt University, Nashville, TN 37240, USA

ARTICLE INFO

Article history:
Received 13 February 2017
Received in revised form 6 December 2017
Accepted 18 December 2017
Available online xxxx
Communicated by D. Perrin

Keywords:
Combinatorics on words
Universal cycles

ABSTRACT

Chen, Kitaev, Mütze, and Sun recently introduced the notion of universal partial words, a generalization of universal words and de Bruijn sequences. Universal partial words allow for a wild-card character $\diamond$, which is a placeholder for any letter in the alphabet. We extend results from the original paper and develop additional proof techniques to study these objects. For non-binary alphabets, we show that universal partial words have periodic $\diamond$ structure and are cyclic, and we give number-theoretic conditions on the existence of universal partial words. In addition, we provide an explicit construction for an infinite family of universal partial words over non-binary alphabets.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Universal cycles of a wide variety of combinatorial structures have been well studied [5]. The best known examples are the *de Bruijn sequences*, cyclic sequences over an alphabet A that contain each word of length n as a substring exactly once. For example, (00010111) is a de Bruijn sequence for $A = \{0, 1\}$ and $n = 3$.

De Bruijn sequences can also be written as non-cyclic sequences. We denote the set of all words of length n over a finite alphabet A by A^n. A *universal word* for A^n is a word w such that each word in A^n appears exactly once as a substring of w. For example, 0001011100, 0010111000, and 1011100010 are universal words for $\{0, 1\}^3$, obtained by splitting the de Bruijn sequence (00010111) and repeating the first $n - 1 = 2$ characters at the end so that the same substrings occur when read non-cyclically. Every universal word corresponds to a de Bruijn sequence in this way.

* Corresponding author.
E-mail addresses: bennet@ku.edu (B. Goeckner), corbin.groothuis@huskers.unl.edu (C. Groothuis), chettle@gatech.edu (C. Hettle), bkell@alumni.cmu.edu (B. Kell), prk213@lehigh.edu (P. Kirkpatrick), rkirsch@huskers.unl.edu (R. Kirsch), ryan.w.solava@vanderbilt.edu (R. Solava).

[1] Research supported in part by National Science Foundation grant # 1604733 “Collaborative Research: Rocky Mountain-Great Plains Graduate Research Workshop in Combinatorics”.
[2] Research supported in part by National Security Agency Grant # H98230-16-1-0018 “The 2016 Rocky Mountain-Great Plains Graduate Research Workshop in Combinatorics”.
[3] Research supported in part by a generous grant from the Institute of Mathematics and its Applications.

https://doi.org/10.1016/j.tcs.2017.12.022
0304-3975/© 2017 Elsevier B.V. All rights reserved.

The length of a universal word for A^n is $|A|^n + n - 1$. It is known that universal words for A^n exist for any n; they can be found through Eulerian and Hamiltonian cycles in the de Bruijn graph [4]. However, as there are $|A|^{|A|^n+n-1}$ words over A of the correct length, a brute-force search for universal words would quickly become intractable.

Now consider extending the alphabet A to $A \cup \{\diamond\}$, where $\diamond \notin A$ is a wild-card character that can correspond to any letter of A. *Partial words* are sequences of characters from $A \cup \{\diamond\}$. Partial words are natural objects in coding theory and theoretical computer science. There are also applications in molecular biology and data communication [2]. For example, when representing DNA and RNA as a string for computing purposes, the $\diamond$ character can take the place of any unknown nucleotide.

For any partial word u, we denote by u_i the ith character of u. Given partial words u and v, we say that $u \subset v$ (or u is a *factor* of v, or v *covers* u), if u can be found as a consecutive substring of v after possibly replacing some $\diamond$ characters in v with a letter from A. Formally, $u \subset v$ if there exists i such that $u_j = v_{i+j}$ for $1 \le j \le |u|$ whenever $v_{i+j} \in A$. For example, $01\diamond1 \subset 1\diamond1\diamond110$, but $01\diamond1 \not\subset 1\diamond10110$.

Chen, Kitaev, Mütze, and Sun [4] introduced the notion of *universal partial words* for A^n, generalizing universal words.

Definition 1.1. A *universal partial word* for A^n is a partial word w that covers each word in A^n exactly once.

For example, $\diamond\diamond0111$ is a universal partial word for $\{0,1\}^3$. Allowing $\diamond$'s decreases the number of characters required to cover all words of length n, so universal partial words may be useful in questions related to storing information in compact form. Trivially, $\diamond^n$ is a universal partial word for A^n for any A and n. A universal partial word w is called *trivial* if all of its characters are diamonds ($w = \diamond^n$) or none are (i.e. w is a universal word for A^n). We are interested in the existence of nontrivial universal partial words.

To begin, we present some preliminaries in Section 2. In Section 3, we generalize the single-diamond study of [4] to single strings of diamonds, and prove the following result:

Theorem 3.4. *For $|A| \ge 2$ and $2 \le k \le n/2$, there does not exist a universal partial word $w = u\diamond^k v$ where u and v are (possibly empty) words.*

Every universal word can be made into a de Bruijn sequence by deleting the last $n-1$ characters, which are the same as the first $n-1$ characters. We call a universal partial word *cyclic* if its first and last $n-1$ characters are the same and non-overlapping, so that deleting the last $n-1$ characters and then reading cyclically yields the same substrings. In this sense, every universal word is cyclic, but some universal partial words over $\{0,1\}$, such as $\diamond\diamond0111$, are not. We prove in Section 4 that all non-binary universal partial words are cyclic.

Theorem 4.8. *If w is a nontrivial, non-binary universal partial word for A^n, then w is cyclic.*

The possible diamond structures of non-binary universal partial words are more limited than when considering the binary case, which both restricts the possible lengths of universal partial words and gives rise to number theoretic conditions limiting their existence.

In Section 5, we show that there are no nontrivial universal partial words for $n \le 3$ over non-binary alphabets. For $n = 4$, there are no universal partial words over alphabets of odd size. However, we show the following:

Theorem 5.2. *For any alphabet A of even size, there exists a nontrivial universal partial word for A^4.*

In Section 5, we give an explicit construction for words of this type. They are the first known nontrivial, non-binary universal partial words. Finally, in Section 6, we discuss some open questions.

2. Preliminaries

An *alphabet* A is a set of symbols, which we call *letters*. A *character* is a letter or $\diamond$. Throughout this paper, we will denote the size of the alphabet A by a and assume without loss of generality that the alphabet is $\{0, 1, \dots, a-1\}$.

A *word* over an alphabet A is a sequence of letters. A *partial word* is a sequence of characters from $A \cup \{\diamond\}$. Note that a word cannot contain a $\diamond$; we sometimes use the term *total word* instead of *word* for emphasis.

Note that if w is a universal partial word for A^n, then the reverse of w is as well. We will refer to this as the *reversal property*. Permuting the letters of A in w also yields a universal partial word.

Given a universal partial word w, a *window* of w is a string of n consecutive characters in w. The *frame* of a partial word is the word over $\{\_, \diamond\}$ obtained by replacing all letters from A by the "_" character. For example, the partial word $xy\diamond z\diamond$ has frame _ _ $\diamond$ _ $\diamond$. A *window frame* is the frame of a window. In the above example, if $n = 3$, the second window is $y\diamond z$, and the second window frame is _ $\diamond$ _.

Theorem 4.1 shows that when $a \ge 3$, every window has the same number of diamonds, so for a universal partial word w over a non-binary alphabet, we define the *diamondicity* of w as the number of diamonds that appear in any window of w. Furthermore, some universal partial words over binary alphabets also have well-defined diamondicity.

Please cite this article in press as: B. Goeckner et al., Universal partial words over non-binary alphabets, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.12.022

Download English Version:

https://daneshyari.com/en/article/6875625

Download Persian Version:

https://daneshyari.com/article/6875625

Daneshyari.com