International Journal of Algebra and Computation
Vol. 15, No. 4 (2005) 717–724

UNAVOIDABLE AND ALMOST UNAVOIDABLE SETS OF WORDS

JASON P. BELL

Department of Mathematics, Simon Fraser University
8888 University Dr., Burnaby, BC V5A 1S6, Canada
belljp@umich.edu

Received 4 August 2004
Revised 4 September 2004

Communicated by D. Perrin

A set of words over a finite alphabet is called an unavoidable set if every word of sufficiently long length must contain some word from this set as a subword. Motivated by a theorem from automata theory, we introduce the notion of an almost unavoidable set and prove certain asymptotic estimates for the size of almost unavoidable sets of uniform length.

Keywords: Unavoidable sets; almost unavoidable sets.

Mathematics Subject Classification: Primary: 68R15; Secondary: 05A99

1. Introduction

Let $\mathcal{X} = \{X_1, \ldots, X_k\}$ be a finite alphabet. We say that a set $\mathcal{S}$ of words is *unavoidable* if every word of sufficiently long length has some element of $\mathcal{S}$ as a subword. We say that $\mathcal{S}$ is n-good if it is unavoidable and every element of $\mathcal{S}$ has length n. Following Schützenberger [6], we define

$$\alpha(k, n) = \inf\{\#\mathcal{S} | \mathcal{S} \text{ is an } n\text{-good set on a } k \text{ letter alphabet}\}. \tag{1.1}$$

Schützenberger proved that

$$\lim_{\max\{k,n\}\to\infty} nk^{-n}\alpha(k, n) = 1, \tag{1.2}$$

where the limit is taken over $k > 1, n \geq 1$. Mykkeltveit [4] proved that

$$\alpha(k, n) = \frac{1}{n}\sum_{d|n} k^{n/d}\phi(d). \tag{1.3}$$

His proof was actually a proof of Golomb's conjecture in a special case, but the formula for $\alpha(k, n)$ was a consequence of this proof. We note that Schützenberger's result given in Eq. (1.2) is an easy consequence of Eq. (1.3). Recently, Champarnaud, Hansel and Perrin [1] gave a direct and clever proof of this formula.

In the next section, we introduce the notion of an *almost unavoidable* set of words, motivated by a dichotomy result from automata theory. We look at the smallest possible size of an almost unavoidable set on a k letter alphabet in which each word has length n and we obtain some Schützenberger-type results for such an object. We conclude with an open question concerning the asymptotic behavior of the size of such a set.

2. Almost Unavoidable Sets

The notion of an unavoidable set of words on a finite alphabet arises quite naturally; it is simply a set of words with the property that any word of sufficiently long length must contain at least one of the words from this set as a subword. In general, given a set of words on a finite alphabet $\mathcal{S}$ one can construct the function

$$f_{\mathcal{S}}(n) = \begin{array}{l}\text{the number of words of length } n \text{ which do not have} \\ \text{an element of } \mathcal{S} \text{ as a subword.}\end{array} \tag{2.4}$$

The following theorem is a well-known result from automata theory (see, for example, Lothaire [3, Chap. 1]).

Theorem 2.1. *Let $\mathcal{X}$ be a regular language. Then the sequence $f(n)$ of words of $\mathcal{X}$ of length n satisfies a linear recurrence relation.*

It is well-known that given a finite alphabet and a finite set of words $\mathcal{S}$, the set of words avoiding $\mathcal{S}$ is regular. Hence

$$\sum_{n=0}^{\infty} f_{\mathcal{S}}(n)x^n$$

is a rational function. Thus either there is some polynomial $p(x)$ such that $f_{\mathcal{S}}(n) \leq p(n)$ for $n \geq 1$, or there is some $C > 1$ such that $f_{\mathcal{S}}(n) > C^n$ for all n sufficiently large.

This theorem says that there is a large "gap" in the possible behavior of the function $f_{\mathcal{S}}$; namely, it either grows exponentially or is bounded by a polynomial. If $\mathcal{S}$ is allowed to be infinite, this dichotomy does not hold. For example, take a two-letter alphabet $\{X, Y\}$ and take

$$\mathcal{S} = \{X^2\} \cup \{XY^iXY^jX | i > j\}.$$

Any word which does not contain an element of $\mathcal{S}$ as a subword is of the form

$$Y^{p_1}XY^{i_1}XY^{i_2}\cdots XY^{i_d}XY^{p_2}, \tag{2.5}$$

where $i_1 \le i_2 \le \cdots \le i_d$. In particular, if $1 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ is a partition of n, then $XY^{\lambda_1}XY^{\lambda_2}\cdots XY^{\lambda_d}X$ is a word of length at most $2n+1$ which does not have an element of $\mathcal{S}$ as a subword. We conclude that

$$\sum_{i=0}^{2n+1} f_{\mathcal{S}}(i) \ge p(n) \sim \frac{1}{4n\sqrt{3}}\exp\bigl(\pi\sqrt{2n/3}\bigr), \tag{2.6}$$

where $p(n)$ is the number of partitions of n. Thus the growth of $f_{\mathcal{S}}$ cannot be bounded by some polynomial.

Conversely, given a word of length n of the form given in item (2.5), we have at most $(n+1)^2$ choices for p_1 and p_2, and the sequence $i_1 \le i_2 \le \cdots \le i_d$ gives a partition of a number that is less than or equal to n. Hence

$$f_{\mathcal{S}}(n) \le (n+1)^2\bigl(p(0)+\cdots+p(n)\bigr) \le (n+1)^3 p(n) \sim \frac{n^2}{4\sqrt{3}}\exp\bigl(\pi\sqrt{2n/3}\bigr).$$

This shows that $f_{\mathcal{S}}$ does not grow exponentially, either.

Motivated by the growth dichotomy, we make the following definition.

Definition 2.2. Let $\mathcal{S}$ be a finite set of words on a finite alphabet $\mathcal{X}$. We say that $\mathcal{S}$ is *almost unavoidable* if $f_{\mathcal{S}}(n) = \mathrm{O}(n^d)$ for some d.

Note that $\mathcal{S}$ is unavoidable if and only if $f_{\mathcal{S}}(n) = 0$ for n sufficiently large.

Remark 2.3. If $|\mathcal{X}| = 1$, then there is at most one word of length n for each n and hence any finite set (including the empty set) is almost unavoidable.

Definition 2.4. We say that a set $\mathcal{S}$ of words of length n is *almost n-good* if $\mathcal{S}$ is almost unavoidable.

Finally, we define

$$\beta(k,n) = \inf\{\#\mathcal{S} | \mathcal{S} \text{ is an almost } n\text{-good set on a } k \text{ letter alphabet}\}. \tag{2.7}$$

We now give a few examples of some almost n-good sets.

Example 2.5. Let $\mathcal{X} = \{X, Y\}$ and let $\mathcal{S} = \{X^2, Y^2\}$. Then $\mathcal{S}$ is almost 2-good.

This is easily seen to be the case, because for any word which avoids $\mathcal{S}$, any occurrence of X must immediately be followed by Y and any occurrence of Y must be immediately followed by X. Hence there are only two words of length m which avoid $\mathcal{X}$; namely, the initial subword of length m of $XYXY\cdots$, and the initial subword of length m of $YXYX\cdots$.

Example 2.6. Let $\mathcal{X} = \{X, Y\}$ and let

$$\mathcal{S} = \{WX^2|\text{length}(W) = n-2\} \cup \{WY^2|\text{length}(W) = n-2\}.$$

Then $\mathcal{S}$ is almost n-good.

This is similar to the preceding example. Note that if W is a word of length $m \geq n$ which avoids $\mathcal{S}$, then after the $(n-2)$nd position of W, there can be no occurrences of X^2 or Y^2. Hence a word of length at least $n-2$ which avoids $\mathcal{S}$ is either of the form $WXYXY\cdots$, or of the form $WYXYXY\cdots$, where W is a word of length $n-2$. Thus for words of length $m \geq n$, there are at most 2^{n-1} words which avoid $\mathcal{S}$ and so $\mathcal{S}$ is almost unavoidable.

Example 2.7. Let $\mathcal{X} = \{X_1, \ldots, X_k\}$ and let $\mathcal{S} = \{X_jX_i|j > i\}$. Then $\mathcal{S}$ is almost 2-good.

In this case, the only words which avoid $\mathcal{S}$ are of the form $X_1^{i_1}X_2^{i_2}\cdots X_k^{i_k}$. An easy counting argument shows that there are $\binom{m+k-1}{k-1}$ such words of length m, which is a polynomial in m of degree $k-1$.

Proposition 2.8. *We have the following values*:

- $\beta(1, n) = 0$ *for* $n \geq 1$;
- $\beta(k, 1) = k-1$ *for* $k \geq 1$; *and*
- $\beta(k, 2) = \binom{k}{2}$ *for* $k \geq 1$.

Proof. From Remark 2.3, we see that $\beta(1, n) = 0$ for all n. If we take a subset $\mathcal{S}$ of a k-letter alphabet of cardinality at most $k-2$ then we must be omitting at least 2 letters, say X_1 and X_2. It is clear then that the 2^n words of length n on X_1 and X_2 will avoid the elements of our set and hence $f_{\mathcal{S}}(n) \geq 2^n$, and so $\mathcal{S}$ cannot possibly be almost 1-good. On the other hand, taking a set of $k-1$ letters leaves only one letter and so a $k-1$ element set of letters is almost 1-good. Thus $\beta(k, 1) = k-1$.

Observe that $\mathcal{S} = \{X_iX_j|1 \leq i < j \leq k\}$ is an almost 2-good set of size $\binom{k}{2}$ and so $\beta(k, 2) \leq \binom{k}{2}$. Next, suppose that $\mathcal{S}$ is almost 2-good. We have two cases.

Case I. $X_i^2 \notin \mathcal{S}$ for some i.

By relabeling if necessary, we may assume that $X_1^2 \notin \mathcal{S}$. We claim that for each j either X_1X_j or X_jX_1 is in $\mathcal{S}$. To see this, suppose that there is some j such that neither X_1X_j nor X_jX_1 is in $\mathcal{S}$. Then the words $V = X_1X_j$ and $W = X_1^2X_j$ have the property that any word in V and W (regarding V and W for the moment as letters) avoids $\mathcal{S}$. Hence there are at least 2^m words of length at most $3m$ which avoid $\mathcal{S}$, contradicting the fact that $\mathcal{S}$ is almost 2-good. Clearly the set $\mathcal{S}_1 = \mathcal{S} \cap \{X_iX_j|1 < i, j \leq k\}$ is almost 2-good for the alphabet $\{X_2, \ldots, X_k\}$ and so by the inductive hypothesis, $\mathcal{S}_1$ has at least $\binom{k-1}{2}$ elements. We have seen that $\mathcal{S}$ has at least $k-1$ additional elements and so the claim follows in this case.

Case II. $X_i^2 \in \mathcal{S}$ for all i.

In this case let i, j be distinct numbers. Let

$$\mathcal{T} = \{(i,j) | 1 \leq i < j \leq k,\ X_iX_j, X_jX_i \notin \mathcal{S}\}.$$

If $\mathcal{T}$ is empty, then $\mathcal{S} \geq \binom{k}{2}$, and so it is no loss of generality to assume that $\mathcal{T}$ is non-empty. Let $(i,j) \in \mathcal{T}$, then $X_iX_\ell X_j$ and $X_jX_\ell X_i$ have some word(s) in $\mathcal{S}$ as subwords for each ℓ, for if, say $X_iX_\ell X_j$ avoids $\mathcal{S}$, then every word on $V = X_iX_\ell X_j$ and $W = X_iX_j$ (again regarding V and W as letters) avoids $\mathcal{S}$; a simple counting argument, similar to the one employed in the preceding case, shows that $\mathcal{S}$ is not almost 2-good. Thus there are at least $2k-3$ words in $\mathcal{S}$ which have either an X_i or an X_j in them somewhere. By the inductive hypothesis, $\mathcal{S}_2 = \mathcal{S} \cap \{X_\ell X_m | \ell, m \notin \{i,j\}\}$ is almost 2-good for a $k-2$ letter alphabet and hence has at least $\binom{k-2}{2}$ elements. Since $\mathcal{S}$ has at least $2k-3$ additional elements, we see that $\mathcal{S}$ has cardinality at least $\binom{k}{2}$. □

To obtain our first estimate for $\beta(k,n)$ for $k \geq 1$, we use a theorem of Golod from ring theory. Given a field F, we let $F\{X_1, \ldots, X_k\}$ denote the free algebra on $X_1, \ldots, X_k$; this can be thought of intuitively as being the ring of noncommutative polynomials on $X_1, \ldots, X_k$.

Theorem 2.9 (Golod). *Let F be a field and let I be a homogeneous ideal in $F\{X_1, \ldots, X_k\}$, generated by a set $\mathcal{S}$ of homogeneous elements, each of degree at least 2. Suppose that $\mathcal{S}$ has at most m_i elements of degree i for each $i \geq 2$. Then if the power series expansion of*

$$G(x) := \left(1 - kx + \sum_{j\geq 2} m_j x^j\right)^{-1}$$

has non-negative coefficients, then the dimension of the vector space spanned by the images of words of length d in $F\{X_1, \ldots, X_k\}/I$ is at least as large as the coefficient of x^d in the power series expansion of $G(x)$.

Proof. See Rowen [5, Lemma 6.2.7, p. 117]. □

To see the connection between this theorem and finding almost n-good sets, note that given an alphabet $\mathcal{X} = \{X_1, \ldots, X_k\}$ and a set of words $\mathcal{S}$, we can regard the elements of $\mathcal{S}$ as being elements of $F\{X_1, \ldots, X_k\}$. We let I denote the ideal in $F\{X_1, \ldots, X_k\}$ generated by the elements of $\mathcal{S}$. Then I is a homogeneous ideal and the dimension of the vector space spanned by the images of words of length i in $F\{X_1, \ldots, X_k\}/I$ is simply $f_{\mathcal{S}}(i)$. In the case that the ideal I is generated by a finite set of words $\mathcal{S}$ on a k-letter alphabet, all of which have length n, this is simply saying that if the power series expansion of

$$1/(1 - kx + (\#\mathcal{S})x^n)$$

has non-negative coefficients, then the coefficient of x^d gives a lower bound on $f_{\mathcal{S}}(d)$ for $d \geq 1$.

Proposition 2.10. *Let*

$$\sum_{i=0}^{\infty} a_i x^i := \left(1 - kx + k^n(1-1/n)^{n-1}n^{-1}x^n\right)^{-1}.$$

Then $a_d \geq (k - k/n)a_{d-1}$ *for each* $d \geq 1$. *In particular,* $a_d \geq (k - k/n)^d$ *for all* $k \geq 0$.

Proof. We have $a_1 = k \geq (k - k/n) = (k - k/n)a_0$. Suppose now that the claim is true for $d < m$. We have a recursion,

$$a_m = ka_{m-1} - k^n a_{m-n}(1-1/n)^{n-1}n^{-1},$$

where a_i is defined to be 0 for $i < 0$. By the inductive hypothesis, $a_{m-1} \geq (k - k/n)^{n-1}a_{m-n}$. Thus

$$\begin{aligned}
a_m &= ka_{m-1} - k^n(1-1/n)^{n-1}n^{-1}a_{m-n} \\
&= (k - k/n)a_{m-1} + kn^{-1}a_{m-1} - k^n(1-1/n)^{n-1}n^{-1}a_{m-n} \\
&\geq (k - k/n)a_{m-1} + \left(k(k-k/n)^{n-1}n^{-1} - k^n(1-1/n)^{n-1}n^{-1}\right)a_{m-n} \\
&= (k - k/n)a_{m-1} + \frac{k^n a_{m-n}}{n}\left((1-1/n)^{n-1} - (1-1/n)^{n-1}\right) \\
&= (k - k/n)a_{m-1}.
\end{aligned}$$

The result now follows by induction. □

Corollary 2.11. *Let* $\mathcal{S}$ *be an almost* n*-good set of words on a* k*-letter alphabet and let* $k > 1$. *Then*

$$\#\mathcal{S} \geq \frac{k^n(1-1/n)^{n-1}}{n}.$$

Proof. By Proposition 2.10 and Theorem 2.9, if $\#\mathcal{S} < k^n(1-1/n)^{n-1}n^{-1}$, then $f_{\mathcal{S}}(m) \geq (k - k/n)^m$ and so if $k > 1, n > 1$ and $(k, n) \neq (2, 2)$, then $f_{\mathcal{S}}(m) > (4/3)^m$ and so $\mathcal{S}$ cannot possibly be almost n-good. If $n = 1$, then $\mathcal{S}$ must have at least $k-1$ elements to be almost n-good, and so the statement is true in this case. Finally, $\beta(2, 2) = 1$ and $k^n(1-1/n)^{n-1}n^{-1} = 1$ when $(k, n) = (2, 2)$, and thus the claim is true in this case. □

We obtain the following result as an immediate corollary.

Corollary 2.12. *For* $k > 1$, *we have* $nk^{-n}\beta(k, n) \geq (1-1/n)^{n-1} \geq 1/e$.

The following proposition gives a better lower bound for $\beta(k, n)$ for (k, n) satisfying $k > 1$ and $n \geq 5$.

Proposition 2.13. *For* $k > 1, n \geq 1$, $\beta(k, n) > k^n/(2n + 2)$.

Proof. Suppose that $\mathcal{S}$ is an almost n-good set on $\mathcal{X} = \{X_1, \ldots, X_k\}$ of size at most $k^n/(2n+2)$ with $k > 1$. Let A and B be two distinct words of length n. We can form $2k$ words of length $2n+1$ by taking words of the form AX_iB and BX_iA for $1 \le i \le k$. We claim that at least k of these words must have an element of $\mathcal{S}$ as a subword; otherwise, we would have (after possibly interchanging A and B) at least one word of the form AX_iB and at least 2 words of the form BX_jA which would have no element of $\mathcal{S}$ as a subword. Suppose that this is the case; let us write these words AX_iB and BX_jA and $BX_\ell A$ and let $W = AX_iBX_j$ and let $V = AX_iBX_\ell$. Observe that any word on W and V (regarding them, for the moment, as letters) will not have any element of $\mathcal{S}$ as a subword, since A and B both have length at least n. Moreover, W and V each have length $2n+2$. Hence there are at least 2^M words of length $(2n+2)M$ which do not have an element of $\mathcal{S}$ as a subword. But this says that $f_{\mathcal{S}}(m) \ge C^m$ for infinitely many values of m with $C = 2^{1/(2n+2)} > 1$, contradicting the fact that $\mathcal{S}$ is almost n-good. There are $\binom{k^n}{2}$ pairs of distinct words of length n and the above argument shows that at least $k\binom{k^n}{2}$ of the words of the form AX_iB with A and B distinct must have an element of $\mathcal{S}$ as a subword.

For a word A of length n, there are at least $k-1$ words of the form AX_iA which have an element of $\mathcal{S}$ as a subword. Otherwise we would have two words AX_iA and AX_jA with no such subwords. Then $V = AX_i$ and $W = AX_j$ would have the property that every word in V and W (regarding them momentarily as letters) would have no elements of $\mathcal{S}$ occurring as subwords and as before we have that $\mathcal{S}$ cannot possibly be almost n-good in this case. Thus there are $k^n(k-1)$ words of the form AX_iA with A of length n which have an element of $\mathcal{S}$ as a subword. Combining these two cases, we see that there are at least

$$k\binom{k^n}{2} + k^n(k-1)$$

words of length $2n+1$ which have an element of $\mathcal{S}$ as a subword.

We can get an upper bound on the number of words of length $2n+1$ with an element of $\mathcal{S}$ as a subword. Observe that in such a word there are $n+1$ possibilities for the spot at which the first letter of the first occurrence of an element of $\mathcal{S}$ can appear. Then there are $\#\mathcal{S}$ choices for the word in $\mathcal{S}$ which appears. Finally, there are k^{n+1} choices for the remaining letters in the word. Hence there are at most $(n+1)(\#\mathcal{S})k^{n+1}$ words of length $2n+1$ with an element of $\mathcal{S}$ as a subword. Hence if $\mathcal{S}$ is almost n-good, we have

$$k\binom{k^n}{2} + k^n(k-1) \le (n+1)(\#\mathcal{S})k^{n+1}.$$

Since $k > 1$, we have

$$\frac{k^n}{2} \le \frac{(k^n-1)}{2} + \frac{k-1}{k} \le (n+1)(\#\mathcal{S}).$$

It follows that $\#\mathcal{S} \ge \frac{k^n}{2n+2}$. □

Putting the estimates from Corollary 2.12 and Proposition 2.13 together we obtain the following lower bound.

$$\max\left\{\frac{n}{2n+2}, (1-1/n)^{n-1}\right\} \leq k^{-n} n\beta(k,n).$$

Using this result along with the facts $\beta(k,1) = k-1$, $\beta(k,n) \leq \alpha(k,n)$, and the result of Schützenberger given in Eq. (1.2), we obtain the following theorem.

Theorem 2.14.

$$\frac{5}{12} \leq \liminf_{\max(k,n)\to\infty} k^{-n} n\beta(k,n) \leq \limsup_{\max(k,n)\to\infty} k^{-n} n\beta(k,n) = 1,$$

where the limits are taken over $k > 1$ *and* $n \geq 1$.

We finish with a conjecture concerning the behavior of $\beta(k,n)$.

Question 2.15. Does the limit of $nk^{-n}\beta(k,n)$ as $\max(k,n) \to \infty$, taken over $k > 1$ and $n \geq 1$, exist?

Note that if the limit exists it must be equal to 1, since $\beta(k,1)k^{-1} = \frac{(k-1)}{k} \to 1$ as $k \to \infty$. If the limit does indeed exist and is equal to 1, it would be interesting to understand the asymptotic behavior of $\alpha(k,n) - \beta(k,n)$.

Acknowledgments

I thank Chris Saker and Dominique Perrin for many helpful comments and suggestions.

References

[1] J. M. Champarnaud, G. Hansel and D. Perrin, Unavoidable sets of constant length, *Int. J. Algebra Comput.* **14**(2) (2004) 241–251.
[2] M. Lothaire, *Combinatorics on Words*, Encyclopedia of Mathematics and Its Applications, Vol. 17 (Addison-Wesley Publishing Co., Reading, MA, 1983).
[3] M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and Its Applications, Vol. 90 (Cambridge University Press, Cambridge, 2002).
[4] J. Mykkeltveit, A proof of Golomb's conjecture for the de Bruijn graph, *J. Combinatorial Theory Ser. B* **13** (1972) 40–45.
[5] L. H. Rowen, Ring Theory. Vol. II, Pure and Applied Mathematics, Vol. 128 (Academic Press, Inc., Boston, MA, 1988).
[6] M. P. Schützenberger, On the certain synchronizing properties of certain prefix codes, *Inform. Comput.* **7** (1964) 23–36.