INFORMATION AND CONTROL 29, 1–10 (1975)

The Dimension of a Formal Language

WILLIAM M. CONNER

Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

A connection is established between formal language theory and mathematical analysis by associating the symbols of strings of a language with the digits of expansions of points in the unit interval. A language is made to correspond to a particular subset of the unit interval, and the dimension of a language is defined as the Hausdorff dimension of this subset. It is shown that the dimension of a language is less than or equal to its channel capacity, and it is shown that a statement involving the dimension of a language can be added to a list of criteria developed by Brainerd and Knode (1972) for determining that a language is not recognizable by a finite automaton.

1. INTRODUCTION

This paper introduces the idea of viewing a formal language as a subset of the unit interval in order that techniques from mathematical analysis may be employed. The particular concept from analysis which will be used here is that of Hausdorff dimension. It is defined as follows (see Billingsley, 1965).

Let $S = \{0, 1,..., b - 1\}$ where $b \geqslant 2$ is an integer, let $x \in (0, 1]$, and let $x = \sum_{i=1}^{\infty} x_i b^{-i}$ be the nonterminating base b expansion of x. Define $b_i(x) = x_i$ for all i, i.e., $b_i(x)$ is the ith digit of the nonterminating base b expansion of x. A set of the form $\{x : b_i(x) = s_i\ ,\ i = 1,..., n\}$, where $s_i \in S$, is denoted $[s_1 ,..., s_n]$ and is called a cylinder of length b^{-n}. Note that $[s_1 ,..., s_n]$ is a half-open (open on the left) b-adic interval of length b^{-n} (i.e., an interval of the form $(j/b^n, (j + 1)/b^n]$ for some j, $0 \leqslant j \leqslant b^n - 1$).

Now let $M \subseteq (0, 1]$, let α and ρ be positive real numbers, and let λ be Lebesgue measure. Define $\lambda_\alpha(M, \rho) = \inf \sum_i \lambda(v_i)^\alpha$, where the infimum is taken over all ρ-coverings of M, a ρ-covering being a covering by cylinders v_i with $\lambda(v_i) < \rho$. It is clear that $\lambda_\alpha(M, \rho) \leqslant \lambda_\alpha(M, \rho')$ for $\rho' < \rho$, so the limit

$$\lambda_\alpha(M) = \lim_{\rho \to 0} \lambda_\alpha(M, \rho)$$

exists (but may be infinite). It can be shown that for fixed M there is an α_0

such that $\lambda_\alpha(M) = \infty$ for $\alpha < \alpha_0$ and $\lambda_\alpha(M) = 0$ for $\alpha > \alpha_0$. The number α_0 is called the (Hausdorff) dimension of M and is denoted by $\dim M$. Some properties of dimension are given in the following proposition (Billingsley, 1965).

PROPOSITION 1. *Dimension has the following properties:*

(i) *The dimension of a set lies between* 0 *and* 1;
(ii) *The dimension of a countable set is* 0;
(iii) *The dimension of a set of positive Lebesgue measure is* 1; *and*
(iv) $\dim (M_1 \cup M_2) = \max (\dim M_1, \dim M_2)$ *for any two subsets* M_1 *and* M_2 *of the unit interval.*

In Section 2 a subset of the unit interval associated with a language is described and the dimension of a language is defined. Section 3 gives examples of the dimension and channel capacity of certain languages. In Section 4 an inequality relationship between the dimension and channel capacity of a language is proven. In Section 5 it is shown that a statement involving the dimension of a language can be added to a list of criteria developed by Brainerd and Knode (1972) for determining that a set is not recognizable by a finite automaton. Finally, Section 6 gives examples using the result of Section 5.

2. The Dimension of a Language

Let Σ be a finite alphabet of cardinality $b \geqslant 2$, let L be a language over Σ (i.e., $L \subseteq \Sigma^*$ where Σ^* is the set of all finite strings of elements from Σ), and let f be a one-to-one onto function from Σ to S. Define a subset $M_L{}^f$ of $(0, 1]$ associated with L and f as follows.

DEFINITION. $M_L{}^f = \{x \in (0, 1]$: given a positive integer p, there exists an integer $q \geqslant p$ such that the string $f^{-1}[b_1(x)] \cdots f^{-1}[b_q(x)] \in L\}$.

$M_L{}^f$ represents the set of infinite strings which have arbitrarily long initial segments belonging to L.

The dimension of a language L, written $\dim L$, is now defined as the number $\dim M_L{}^f$. We omit reference to f when writing $\dim L$ because of the following proposition.

PROPOSITION 2. *Let f and g be one-to-one onto functions from Σ to S. Then* $\dim M_L{}^f = \dim M_L{}^g$ *for any language L.*

Proof. It is clear from properties (ii) and (iv) of Proposition 1 that the dimension of a set is not changed by the addition or deletion of a countable number of points. In what follows, we will assume that all rational points have been removed from M_L^f and M_L^g. Since the set of rationals is countable, the dimensions of M_L^f and M_L^g will remain unchanged.

Let $\{v_i : i = 1, 2,...\}$ be a covering of M_L^f where each v_i is a cylinder $[s_1^i, s_2^i,..., s_{n_i}^i]$. Let u_i be the cylinder

$$[g(f^{-1}(s_1^i)),..., g(f^{-1}(s_{n_i}^i))];$$

we show now that $\{u_i : i = 1, 2,...\}$ is a covering of M_L^g. Let $y \in M_L^g$ and let $y_i = b_i(y)$ for all $i \geqslant 1$. Then it is clear that

$$y' = .f[g^{-1}(y_1)]\,f[g^{-1}(y_2)] \cdots \in M_L^f$$

and thus $y' \in v_k$ for some k. (It is at this point that we wish to avoid consideration of certain rational points. If y was a rational whose expansion was nonterminating in some digit, then it is possible that the expansion above for y' could be the terminating expansion, and y' may not belong to M_L^f.) Now v_k is a cylinder

$$[f(g^{-1}(y_1)),..., f(g^{-1}(y_{n_k}))]$$

so

$$\begin{aligned} u_k &= [g(f^{-1}(f(g^{-1}(y_1)))),..., g(f^{-1}(f(g^{-1}(y_{n_k}))))] \\ &= [y_1 ,..., y_{n_k}], \end{aligned}$$

which covers y. Hence every point of M_L^g belongs to some u_k so $\{u_i : i = 1, 2,...\}$ is a covering of M_L^g.

We have shown that for every cover of M_L^f by a set of cylinders there is another set of the same number of cylinders with the same lengths which covers M_L^g. Similarly it can be shown that every covering of M_L^g has a corresponding covering of M_L^f. This correspondence of coverings of M_L^g and M_L^f is sufficient to show that $\dim M_L^f = \dim M_L^g$. ∎

As a result of Proposition 2, we see that for defining the dimension of a language, there is no loss in generality in assuming that the alphabet Σ is simply S, and that $L \subseteq S^*$. We shall then omit reference to f and refer to simply M_L.

The following proposition demonstrates two properties of the dimension of a language which follow easily from properties of dimension in the unit interval.

PROPOSITION 3. *Let L_1 and L_2 be languages over S.*

(i) $\dim (L_1 \cup L_2) = \max (\dim L_1, \dim L_2)$;

(ii) *If* $L_1 \subseteq L_2$ *then* $\dim L_1 \leqslant \dim L_2$.

Proof. (i) We show first that $M_{L_1 \cup L_2} = M_{L_1} \cup M_{L_2}$. Let $x \in M_{L_1 \cup L_2}$. Then for every positive integer p, there exists an integer $q \geqslant p$ such that $b_1(x) \cdots b_q(x) \in L_1 \cup L_2$. If $x \in M_{L_1}$ then $x \in M_{L_1} \cup M_{L_2}$. If $x \notin M_{L_1}$, then there exists a positive integer r such that for no $s \geqslant r$ does $b_1(x) \cdots b_s(x) \in L_1$. But since $b_1(x) \cdots b_q(x) \in L_1 \cup L_2$ for arbitrarily large q, it must be the case that $b_1(x) \cdots b_q(x) \in L_2$ for arbitrarily large q. Hence $x \in M_{L_2}$ so $x \in M_{L_1} \cup M_{L_2}$.

Now let $x \in M_{L_1} \cup M_{L_2}$. If $x \in M_{L_1}$ then for every positive p, there exists an integer $q \geqslant p$ such that

$$b_1(x) \cdots b_q(x) \in L_1 \subseteq L_1 \cup L_2,$$

so $x \in M_{L_1 \cup L_2}$. A similar argument holds if $x \in M_{L_2}$. Thus $M_{L_1 \cup L_2} = M_{L_1} \cup M_{L_2}$, so by property (iv) of Proposition 1 we have $\dim M_{L_1 \cup L_2} = \max (\dim M_{L_1}, \dim M_{L_2})$.

(ii) Using part (i) and the fact that $L_1 \subseteq L_2$, we have $\dim L_2 = \dim (L_1 \cup L_2) = \max (\dim L_1, \dim L_2) \geqslant \dim L_1$. ▌

Next we prove a proposition that illustrates the structure of the set M_L. Let $N(n)$ be the number of strings of length n in L, and for each such string $s_1{}^i \cdots s_n{}^i$ of length n define $A_n{}^i$ to be the cylinder $[s_1{}^i, \ldots, s_n{}^i]$. Then let

$$A_n \equiv \bigcup_{i=1}^{N(n)} A_n{}^i$$

(where we take $A_n \equiv \varnothing$ if $N(n) = 0$).

PROPOSITION 4. $M_L = \limsup A_n$.

Proof. Let $x \in M_L$ and let p be a positive integer. Then there exists an integer $q \geqslant p$ such that $b_1(x) \cdots b_q(x) \in L$, and hence $x \in A_q$. Therefore $x \in \bigcup_{n=p}^{\infty} A_n$, and since p was arbitrary, $x \in \bigcap_{p=1}^{\infty} \bigcup_{n=p}^{\infty} A_n$, i.e., $x \in \limsup A_n$.

Now let $x \in \bigcap_{p=1}^{\infty} \bigcup_{n=p}^{\infty} A_n$ and let r be a positive integer. Then $x \in \bigcup_{n=r}^{\infty} A_n$ so there exists a $q \geqslant r$ such that $x \in A_q$. Then $b_1(x) \cdots b_q(x) \in L$ so $x \in M_L$. ▌

3. Examples

Dimension, in a sense, measures the size of a language. Another quantity used to measure language size, the (channel) capacity, has been considered by several authors (see, for example, Kuich, 1970). It is defined as follows. Let R be the radius of convergence of the power series $\sum_{n=1}^{\infty} N(n)\, z^n$; the capacity C of a language is defined as $C \equiv \log(1/R)$. By Hadamard's formula,

$$1/R = \limsup_{n\to\infty} [N(n)]^{1/n},$$

and thus

$$C = \limsup_{n\to\infty} [\log N(n)/n].$$

All logs are taken to the base b (the cardinality of the alphabet), and log 0 is taken to be 0. In this section we compute the dimension and capacity of some sample languages; in the next section an inequality relationship between dimension and capacity is proven.

Example 1. Let $L \equiv S^*$. Then $M_L = (0, 1]$ so $\dim L = 1$. Also, $C = \limsup_{n\to\infty} (\log b^n/n) = 1$.

Example 2. Let L be finite. Then $M_L = \varnothing$ so $\dim L = 0$. Also, $C = 0$.

Example 3. Let $S \equiv \{0, 1\}$ and let $L \equiv \{w \in S^*: w$ does not contain two consecutive 1's$\}$. The recurrence relation $N(n) = N(n-1) + N(n-2)$, $n \geqslant 3$, holds, since all strings of length n are accounted for by adding a 0 to the strings of length $n - 1$, or by adding 01 to the strings of length $n - 2$. This is the Fibonacci recurrence relation and it has the solution

$$N(n) = (1/5^{1/2})\{[(1 + 5^{1/2})/2]^{n+2} - [(1 - 5^{1/2})/2]^{n+2}\}.$$

A theorem of the author (1970, Theorem 3.5) can be used to show that $\dim L = C$, and, from the above expression for $N(n)$, this value is seen to be $\log[(1 + 5^{1/2})/2]$.

Example 4. Let $S \equiv \{0, 1, 2\}$ and let $L \equiv \{w \in S^* : w$ does not contain a 1$\}$. Then M_L is the Cantor set and it is well-known that $\dim M_L = \log 2$. Also, $C = \limsup_{n\to\infty} \log 2^n/n = \log 2$.

Example 5. Let $L \equiv \{0^n w : w \in S^*$ and $|w| = n,\ n = 1, 2, \ldots\}$, where $|w|$ denotes the length of w. Then $M_L = \varnothing$ so $\dim L = 0$. Since $N(n) = b^{n/2}$ for n even and $N(n) = 0$ for n odd, $C = 1/2$.

4. Relationship Between Dimension and Capacity

Theorem 1. *Let L be a language with capacity C. Then* $\dim L \leqslant C$.

Proof. By Proposition 4, $M_L = \bigcap_{p=1}^{\infty} \bigcup_{n=p}^{\infty} A_n$. Hence, for any integer $p \geqslant 1$, $\bigcup_{n=p}^{\infty} A_n$ contains M_L, and thus the collection of cylinders

$$\{A_n{}^i : n = p, p+1,...; i = 1, 2,..., N(n)\}$$

is a covering of M_L. Let $\epsilon > 0$ and $\rho > 0$ be given and choose p such that $b^{-p} < \rho$. Then

$$\lambda_{C+\epsilon}(M_L, \rho) \leqslant \sum_{n=p}^{\infty} N(n)\, b^{-(C+\epsilon)n}.$$

Now

$$b^{-(C+\epsilon)} < b^{-C} = b^{-\log 1/R} = R,$$

where R is the radius of convergence of the power series $\sum_{n=1}^{\infty} N(n)\, z^n$. Thus the series $\sum_{n=1}^{\infty} N(n)\, b^{-(C+\epsilon)n}$ converges; call its value V. Then

$$\lambda_{C+\epsilon}(M_L, \rho) \leqslant V < \infty,$$

and since ρ was arbitrary, it follows that $\lambda_{C+\epsilon}(M_L) \leqslant V < \infty$. Thus

$$\dim M_L \leqslant C + \epsilon,$$

and since ϵ was arbitrary, we have $\dim M_L \leqslant C$. ▌

The examples of the preceding section show that the relationship in the theorem cannot be improved to either equality or strict inequality.

5. A New Criterion for Recognizability

We state here for reference purposes a theorem of Brainerd and Knode (1972).

Theorem 2 (Brainerd and Knode). *Let $L \subseteq S^*$, where S is of cardinality b, be recognized by a finite automaton. Let Q be the states of the minimal automaton A which recognizes L. Let $N(n)$ be the number of strings of length n in L, R the radius of convergence of $\sum_{n=1}^{\infty} N(n)\, z^n$, C the capacity of L, and define $\mu(L) \equiv \lim_{m\to\infty} (1/m) \sum_{n=1}^{m} [N(n)/b^n]$. Then the following are equivalent (x, y, and z denote strings in S^*):*

(a) *A has a dead state d accessible from each state $q \in Q$.*

(a′) $\forall q \exists x(qx = d)$, *where qx is the state of A after starting in state q and reading the input x.*

(a″) $\forall y \exists x \forall z(yxz \notin L)$.

(b) $\exists x \forall y \forall z(yxz \notin L)$.

(b′) $\exists x \forall q(qx = d)$.

(b″) $\exists x(S^*xS^* \cap L = \varnothing)$.

(c) $R > (1/b)$, *or equivalently*, $C < 1$.

(d) $N(n)/b^n \to 0$ *as* $n \to \infty$.

(e) $\mu(L) = 0$.

(f) *There is no subsequence of* $\{N(n)\}$ *of the form* $\{N(nt + m)\}_{n=0}^{\infty}$, $t > m \geqslant 0$, *such that* $\lim_{n\to\infty} \{N[(n + 1)\, t + m]/N(nt + m)\} = b^t$.

Another statement, involving the dimension of L, is now added to this list.

THEOREM 3. *The statement below can be added to the list of equivalent statements in Theorem* 2:

$$(g)\ \dim L < 1.$$

Proof. We shall prove that (c) implies (g) and (g) implies (d).

(c) implies (g): By Theorem 1, $\dim L \leqslant C$, and hence $C < 1$ implies $\dim L < 1$.

(g) implies (d): $\dim L < 1$ implies $\lambda_1(M_L) = 0$ and thus $\lambda_1(M_L, \rho) = 0$ for all ρ. Let $\epsilon > 0$ be given. Then there exists a covering $\{v_i : i = 1, 2,...\}$ of $M_L = \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{\infty} A_j$ such that $\sum_{i=1}^{\infty} \lambda(v_i) < \epsilon/2$. Let $E_n \equiv \bigcup_{j=n}^{\infty} A_j$ and let $v \equiv \bigcup_{i=1}^{\infty} v_i$. Suppose $\lambda(E_n - v) \geqslant \epsilon/2$ for all $n \geqslant 1$. $\{E_n - v\}$ is a decreasing sequence of sets and by a well-known result in analysis,

$$\lambda\left[\bigcap_{n=1}^{\infty} (E_n - v)\right] = \lim_{n\to\infty} \lambda(E_n - v) \geqslant \epsilon/2.$$

But

$$\lambda\left[\bigcap_{n=1}^{\infty} (E_n - v)\right] = \lambda\left[\left(\bigcap_{n=1}^{\infty} E_n\right) - v\right] = \lambda(M_L - v) = \lambda(\varnothing) = 0.$$

This contradiction shows that there must exist a positive integer m such that

$\lambda(E_m - v) < \epsilon/2$. Then since $\{E_n - v\}$ is decreasing, $\lambda(E_n - v) < \epsilon/2$ for all $n \geqslant m$. Now

$$\begin{aligned} N(n)\, b^{-n} &= \lambda(A_n) \leqslant \lambda(E_n) \\ &= \lambda(E_n \cap v) + \lambda(E_n - v) \leqslant \lambda(v) + \lambda(E_n - v) < \epsilon \end{aligned}$$

for all $n \geqslant m$. Thus $N(n)\, b^{-n} \to 0$ as $n \to \infty$. ▮

The condition $\dim L < 1$ means intuitively that L gives rise to a "small" set of infinite strings, since $\dim L < 1$ means that M_L has Lebesgue measure 0 (property (iii) of Proposition 1).

6. Examples

A language which satisfies one of the conditions of Theorems 2 and 3 but fails to satisfy another of the conditions is not recognizable by a finite automaton. Four examples will now be given.

Example 6. Let $L = \{ww\colon w \in S^*\}$. Then the capacity $C = 1/2$ so $\dim L \leqslant 1/2$. Thus conditions (c) and (g) are satisfied but (b″) is not, since for each w, $ww\epsilon \in L$. Hence L is not recognizable.

Example 7. The language L of Example 5 is not recognizable since $\dim L = \dim \varnothing = 0$, and since for each w, $0^{|w|}w\epsilon \in L$, i.e., condition (g) is satisfied but (b″) is not.

It is well known that a language A is recognizable if and only if its reversal A^R is recognizable. This fact is useful when applying condition (g) to a language such as L, since computation of $\dim L$ is trivial but computation of $\dim L^R$ is difficult. Hence, both L and L^R should be examined to determine if one gives an easier computation of the dimension.

This example can be slightly generalized. Let f be an unbounded nonnegative integer-valued function defined on the positive integers. Then the language

$$\{0^n w\colon w \in S^* \text{ and } |\, w \,| = f(n),\ n = 1, 2,\ldots\}$$

is not recognizable, since, as before, condition (g) is satisfied but (b″) is not.

Example 8. Let $S = \{0, 1,\ldots, b-1\}$, $b \geqslant 2$, let p_i, $0 \leqslant i \leqslant b-1$, be positive rational numbers such that $\sum_{i=0}^{b-1} p_i = 1$ and $p_i \neq 1/b$ for some i, and let $L \equiv \{w \in S^*: w$ contains i in the proportion p_i, $0 \leqslant i \leqslant b-1\}$. For each $x \in (0, 1]$ define $a_i(x, n)$, $0 \leqslant i \leqslant b-1$, as the number of occurrences of i among $b_1(x),..., b_n(x)$. Now for each $x \in M_L$ the sequence $\{[a_i(x, n)]/n\}_{n=1}^{\infty}$ has a cluster point of p_i, $0 \leqslant i \leqslant b-1$, since $[a_i(x, n)]/n$ has the value p_i for infinitely many n. Thus M_L is contained in the set

$$S \equiv \{x \in (0, 1] : \limsup_{n\to\infty}[a_i(x, n)]/n \geqslant p_i, 0 \leqslant i \leqslant b-1\}.$$

Now let $H \equiv -\sum_{i=0}^{b-1} p_i \log p_i$ and let

$$T \equiv \left\{x \in (0, 1] : \liminf_{n\to\infty} \left(-\frac{1}{n}\sum_{i=0}^{b-1} a_i(x, n) \log p_i\right) \leqslant H\right\}.$$

A theorem of Billingsley (1961, Theorem 2.1) can be used to show that $\dim T \leqslant H$. It is clear that $S \subseteq T$; therefore

$$\dim L \equiv \dim M_L \leqslant \dim S \leqslant \dim T \leqslant H,$$

and $H < 1$ since $p_i \neq 1/b$ for some i (see, for example, Ash (1965, Theorem 1.4.2)). Thus condition (g) is satisfied. However, (b″) is not satisfied since for each $w \in S^*$,

$$0^{n_1}1^{n_2} \cdots (b-1)^{n_b}\, w\epsilon \in L$$

for appropriate integers $n_i \geqslant 0$. Thus L is not recognizable. For example, this technique shows that the language

$$\{w \in \{0, 1\}^*: w \text{ consists of twice as many 0's as 1's}\}$$

is not recognizable.

Example 9. Let $S \equiv \{0, 1,..., b-1\}$, $b \geqslant 2$, and let P_1, P_2, P_3,... be a sequence of strings over S such that (i) a particular string appears only a finite number of times in the sequence, and (ii) $|P_i| \leqslant |P_j|$ implies that P_i is a prefix of P_j (denoted $P_i \mid P_j$). Also, let Q_1, Q_2, Q_3,..., be any enumeration of the strings of S^*, and let $L \equiv \{P_iQ_i : i = 1, 2, 3,...\}$. Because of conditions (i) and (ii), it is clear that there exists a unique infinite string P such that $P_i \mid P$ for all i, and it is easily seen that M_L is either empty (in case the point $.P$ is in the form of a terminating expansion) or consists of the single point $.P$ (in case $.P$ is in the form of a nonterminating expansion). In either case, $\dim L \equiv \dim M_L = 0$, so condition (g) is satisfied. However, (b″) is not satisfied since for each $w \in S^*$, $P_i w\epsilon \in L$, where i is such that $w = Q_i$. Thus L is not recognizable. For this example, conditions (g) and (b″) are particularly easy to use in relation to other techniques to show that L is not recognizable.

If L is any of the languages in the four examples above, then $L \cup K$ is not recognizable for any language K with dim $K < 1$. The reason is that condition (g) is still satisfied and (b″) is still not satisfied.

Acknowledgment

The author is grateful to Dr. Stanley Walljasper for his helpful comments.

Received: July 18, 1973; revised: January 1975

References

Ash, R. (1965), "Information Theory," p. 17, Wiley, New York.

Billingsley, P. (1961), Hausdorff dimension in probability theory II, *Illinois J. Math.* **5**, 291–298.

Billingsley, P. (1965), "Ergodic Theory and Information," pp. 136–145, Wiley, New York.

Brainerd, W. S. and Knode, R. B. (1972), Some criteria for determining recognizability of a set, *Inform. Contr.* **21**, 171–184.

Conner, W. M. (1970), The capacity and ambiguity of a transducer, *Ann. Math. Statist.* **41**, 2093–2104.

Kuich, W. (1970), On the entropy of context-free languages, *Inform. Contr.* **16**, 173–200.