ON THE ENTROPY OF A FORMAL LANGUAGE

by

A. de Luca

Laboratorio di Cibernetica del C.N.R., Arco Felice, Napoli.

and

Istituto di Scienze dell'Informazione dell'Università di Salerno.

0. Introduction

The problem of transmission of information over a communication channel can be studied in two ways which are conceptually different. The first was originated by the fundamental work of C.E. Shannon [16]. The source of information is a probabilistic ergodic source. It satisfies the following property [8] (E-property) which is very important in information theory: the sequences of large length n generated by an ergodic source of entropy H(S) can be divided in two groups. The sequences of the first group, called also standard, have a probability close to $2^{-H(S)n}$ and are in number approximately equal to $2^{H(S)n}$. The remaining sequences of length n have a total probability which vanishes for n diverging. The E-property makes it possible then, to separate from the initial language a sub-language formed by the "high-probability group" of the approximately equiprobable standard sequences, which play the essential role in the coding of Shannon's theorems.

Another approach to information theory that we call "linguistic" exists. This in some respects equivalent to the previous one, consists in considering directly a language L described by the so-called stucture-function f of Mandelbrot [12]. For all n, f(n) gives the number of distinct words of length n contained in the language L. The entropy H(L) of the language L can be, then, defined in a purely combinatorial fashion, as the $\lim_{n\to\infty} (1/n) \ln_2 f(n)$ or, more generally, as $\lim_{n\to\infty} \sup (1/n) \ln_2 f(n)$. It can be regarded as the number of [or the least upper bound to] bits per symbol required, on the average, to specify a word of the language. This concept of entropy was initially introduced by Shannon himself [16] in the case of languages consisting of all "messages" generated by a finite-state communication channel, and was named channel-capacity. This definition of entropy was successively extended by Chomsky and Miller [3] to any finite-state lan-

guage and by other authors [1,10,17] to wider classes of formal languages.The main problem considered by these authors was,essentially,that of giving for some classes of formal languages (as,for instance,some subclasses of context-free languages)general computation methods of the entropy of a language of the class by knowing an underlying grammar generating it.Recently Kaminger [7] proved that there cannot exist a general computation method of the entropy of languages generated by context-sensitive grammars.

The aim of this paper is that of making a preliminary analysis of the problem of transmission of information in the context of the linguistic approach.We shall consider sources of information generating,in the general case,recursively enumerable (r.e.)languages (which can be always produced by type-0 grammars [14])initially described by the structure function only.We are not much interested in the underlying grammar generating a language L,but exclusively in its entropy H(L).The main problem that we shall consider is the one of efficient coding for the words of L relative to a given "effective-decoder" of it.This problem can be faced in a natural way by making use of the Kolmogorov program-complexity [2,9,11,15].In fact the program-complexity $K_{\psi}(\xi)$ of a word ξ,relative the partial recursive (p.r.)function (or decoder) $\psi : Y^{*} \to X^{*}$,can be regarded as the minimal length of a "code-word" of ξ in a "communication-schema" where the "receiver" is an algorithm computing ψ and a "code-word" of a string is a "computer-program" for it (*). Moreover the program-complexity $K(\xi)$ relative to a universal p.r.function (or universal decoder) of the words of L gives a measure of their structural-complexity,since $K(\xi)$ represents the minimum number of bits,to within an additive constant,required to define ξ in an effective manner.The function K allows, therefore,an analysis of the structure of a r.e.language L deeper than those obtained by means of the structure function f.However,as we shall see in the following, the overwhelming majority of the words of L of large length $|\xi|$ has a "compression-coefficient" $\mu(\xi) := K(\xi)/|\xi|$ approximately equal to $(1/|\xi|)\ln_2 f(|\xi|)$ (in the case of a binary code-alphabet).The entropy H(L) can be,then, redefined in terms of Kolmogorov's complexity of the words of L.

1.Entropy of a formal language and Kolmogorov complexity

Let X be a finite (non empty) alphabet of cardinality $\|X\|$,and X^{*} the free-monoid generated by X,that is the set of all finite sequences or words ξ of symbols of X including the empty word λ .The length of a word ξ will be denoted by $|\xi|$ .A language

(*) A generalization of this schema in which the "receiver" is a formal system has been proposed by the author [4-6]

L over the alphabet X is any subset of X^*. Let X^n be the set of all the words over X of length n. For any language L we denote by L_n its subset $L_n := L \cap X^n$. The entropy H(L) of L is the quantity $H(L) := \lim_{n\to\infty} \sup (1/n) \ln_2 f(n)$, where f is the structure-function of L defined as $f(n) := \| L_n \|$, for all n. From the definition one has that H(L) is finite iff L is an infinite language. In this latter case $0 \leq H(L) \leq \ln_2 \| X \|$.

In the following we shall mainly consider recursively enumerable languages. A language L is recursive iff L and its complement $\sim L$ are recursively enumerable [14]. Moreover a r.e. language is recursive iff its structure function is computable. We want now describe a r.e. language in terms of the Kolmogorov program-complexity of its words. We recall that for any p.r.function $\psi : Y^* \to X^*$, the program-complexity $K_\psi(\xi)$ relative to ψ is defined as $K_\psi(\xi) := \min \{ |p| \mid \psi(p) = \xi \}$, where, conventionally $\min \emptyset = +\infty$. The quantity $K_\psi(\xi)$ depends in an essential way on the p.r.function ψ. However a basic theorem due to Solomonoff [18] and Kolmogorov [9] shows that there are asymptotically optimal p.r.functions with respect to which to evaluate the program-complexity. More precisely, a universal p.r.function $U : Y^* \to X^*$ exists with the property that for any other p.r.function $\psi : Y^* \to X^*$ one has that $K_U(\xi) \leq K_\psi(\xi) + c_{U,\psi}$ with $\xi \in X^*$ and $c_{U,\psi} \in N$ (N is the set of nonnegative integers). For any such two universal p.r.functions U_1 and U_2, $|K_{U_1}(\xi) - K_{U_2}(\xi)| \leq$ cost. for all $\xi \in X^*$. Therefore, for all $\xi \in X^*$, $K_{U_1}(\xi)$ and $K_{U_2}(\xi)$ are equal to within an additive constant which can be neglected for high values of the complexity. The program-complexity of a string $\xi \in X^*$ relative to a fixed universal p.r.function U, will be simply denoted by $K(\xi)$. The following theorem, that generalizes a result of Kolmogorov and Martin-Lof [13], shows the relationship existing between the program-complexity of the words of a r.e. language and its structure function.

Theorem 1.1. For all n, such that $L_n \neq \emptyset$, one has that

i. $K(\xi) \leq \ln_d f(n) + O(\ln_d n)$, for all $\xi \in L_n$, where $d = \| Y \|$ and $O(\ln_d n)$ denotes a quantity of the order of $\ln_d n$ when n diverges.

ii. The number of words of L_n for which $K(\xi) \geq \lfloor \ln_d f(n) \rfloor - \delta$, with $\delta \in N$ ($\lfloor x \rfloor$ is the greatest integer $\leq x$) is greater than $f(n)(1 - d^{-\delta}/d - 1)$.

iii. There is a lower bound to the number of words of L_n for which $K(\xi) < \lfloor \ln_d f(n) \rfloor - \delta$ given by $f(n)\, d^{-\delta} - c/d^3 (d-1)((d-1)n + d)^2 - 1/d - 1$, with $c \in N$.

For any r.e. language L the elements of the sublanguage $V(\delta) = \{ \xi \in L \mid K(\xi) \geq \lfloor \ln_d f(n) \rfloor - \delta \}$, whose entropy equals H(L), are called the (f,δ)-random elements of L. With the only exception when $H(L) = \ln_2 \| X \|$ the (f,δ)-random elements of L_n are, for large n, a very small fraction of the set of all sequences of length n. A consequence of theorem 1.1

is that if $L_n \neq \emptyset$ then $V_n(\delta) \neq \emptyset$ since there is at least a word $\xi \in L_n$ such that $K(\xi) \geq \lfloor \ln_d f(n) \rfloor$. Furthermore, when $H(L) > 0$ a p.r.function $r: N \to X^*$ cannot exist such that $r(n) \in V_n(\delta)$ for all n for which $L_n \neq \emptyset$. From this it is easy to derive that if L is recursive and $H(L) > 0$, then $V(\delta)$ cannot be recursively enumerable and K is not computable in L.

2. Effective coding

Let $\psi : Y^* \to X^*$ be a p.r. function. A word $p \in Y^*$ such that $\psi(p) = \xi$ with $\xi \in X^*$ can be regarded as a code-word (or coding), in the alphabet Y, of ξ relative to ψ. The function ψ will be referred to as an effective-decoder (e.d.) for any language $L \subseteq \text{Range } \psi$.

Definition 2.1 . For any given language L ($L \subseteq X^*$) an [effective]-decoder of L is any [p.r.]-function $\psi : Y^* \to X^*$, such that Range $\psi \supseteq L$.

The alphabet Y is called the code-alphabet (Y can be equal to X). For all $\xi \in L$ the inverse-image $\psi^{-1}(\xi) \subseteq Y^*$ is formed by all code-words of ξ. The quantity $C(\psi) := H(\text{Range } \psi)$ will be named the capacity of ψ. For any [effective]-decoder ψ, $C(\psi)$ equals the maximum of the entropy of any [r.e.] language contained in Range ψ.

Definition 2.2 . For any given [effective]-decoder ψ of L an [effective]-encoder of L, relative to ψ, is any [p.r.]-function $\psi_{-1} : X^* \to Y^*$ such that Dom $\psi_{-1} \supseteq L$ and $\psi_{-1}(\xi) \in \psi^{-1}(\xi)$, for all $\xi \in L$.

It is easy to derive, from recursive function theory, the following:

Lemma 2.1 . Given an arbitrary r.e. language L and an e.d. ψ of it there exists always an effective encoder ψ_{-1} of L.

For any partial function $\rho: X^+ \to R$ ($X^+ = XX^*$ and R is the set of real numbers) such that Dom $\rho \supseteq L - \{\lambda\}$, let us denote by $\rho(L)$, $\langle\rho\rangle(n)$ and $\langle\rho\rangle(L)$ respectively the quantities $\rho(L) := \lim_{n\to\infty} \sup\{\rho(\xi) \mid \xi \in L \,\&\, |\xi| \geq n\}$, $\langle\rho\rangle(n) := \sum_{\xi \in L_n} \rho(\xi)/f(n)$ for $L_n \neq \emptyset$, $\langle\rho\rangle(L) := \lim_{n\to\infty} \sup \langle\rho\rangle(n)$.

With respect to any encoder ψ_{-1}, relative to the decoder ψ of L the compression-coefficient $\Psi_{-1}(\xi)$ of a nonempty string ξ of L is defined as $\Psi_{-1}(\xi) := |\psi_{-1}(\xi)|/|\xi|$. Furthermore $\Psi_{-1}(L)$, $\langle\Psi_{-1}\rangle(n)$ and $\langle\Psi_{-1}\rangle(L)$ will be called, respectively, the compression-coefficient of L, the average compression-coefficient of L_n ($L_n \neq \emptyset$) and the average compression-coefficient of L. For any effective decoder ψ of a r.e. language L the quantity $\mu_\psi(\xi) := K_\psi(\xi)/|\xi|$, where ξ is a nonempty word of L and $K_\psi(\xi)$ the program-complexity of ξ relative to ψ, represents the minimal value of the compression coefficient of $\xi \in L$

with respect to all encoders ψ_1 of L relative to ψ. Moreover, one has that $\Psi_{-1}(L) \geq \mu_\psi(L)$, $<\Psi_{-1}>(n) \geq <\mu_\psi>(n)$ and $<\Psi_{-1}>(L) \geq <\mu_\psi>(L)$.

An encoder ψ_{-1} such that $|\psi_{-1}(\xi)| = K_\psi(\xi)$ in L is called absolutely-optimal. Such an encoder is effective iff K_ψ is computable in L. Therefore, the existence of an effective absolutely-optimal encoder of a r.e. language depends in an essential way on the effective decoder ψ. For an infinite r.e. language L there exists always a recursive injection $\psi^0 : Y^* \to X^*$ such that $L = \text{Range } \psi^0$. An effective encoder ψ^0_{-1} (which certainly exists by Lemma 2.1) relative to ψ^0 is absolutely optimal since $\{\psi^0_{-1}(\xi)\} = (\psi^0)^{-1}(\xi)$. On the contrary, from what we said at the end of the previous section, it follows that an absolutely-optimal effective encoder of a recursive language L, with $H(L) > 0$, relative to a universal e.d. U does not exist.

It is easy to prove that the e.d. U, whose capacity equals $\ln_2 \| X \|$, is such that any word $\xi \in X^*$ has an infinite number of code-words. This fact justifies our definition of [effective]-decoder which is a more general one than the usual. Furthermore the basic Solomonoff-Kolmogorov theorem can be restated, in terms of compression-coefficients, in the following form: there exists an effective decoder U of any r.e. language L (that is U is a universal e.d.) which is asymptotically-optimal with respect to all effective-decoders ψ of L, in the sense that for any $\varepsilon > 0$, $\mu_U(\xi) \leq \mu_\psi(\xi) + \varepsilon$, when $\xi \in L_n$ with n sufficiently large. It follows that $\mu_U(L) \leq \mu_\psi(L)$ and $\mu_{U_1}(L) = \mu_{U_2}(L)$ for any such two universal decoders U_1 and U_2. Therefore the quantity $\mu(L) := \mu_U(L)$, which depends only on the r.e. language L, represents the minimal compression-coefficient of L with respect to all e.d. of it.

A corollary of theorem 2.1 of the previous section is the following proposition concerning the compression-coefficient $\mu(\xi) := K(\xi)/|\xi|$ of the words of an infinite r.e. language L.

Proposition 2.1. Given any $\varepsilon > 0$, the (f,δ)-random elements of L of length n, for a fixed δ, are such that $|\mu(\xi) - (1/n)\ln_d f(n)| < \varepsilon$, when n diverges. The fraction of the remaining words of L_n, for which $\mu(\xi) \leq (1/n)\ln_d f(n) - \varepsilon$, becomes as small as one wishes for a sufficiently large δ.

If there exists $\lim_{n\to\infty} (1/n)\ln_2 f(n) = H(L)$ one has that $|\mu(\xi) - H(L)/\ln_2 d| < \varepsilon$ for $\xi \in V_n(\delta)$ when n diverges. A consequence of proposition 2.1 and of theorem 1.1 is that $<\mu>(L) = \mu(L) = H(L)/\ln_2 d$. Furthermore, for any r.e. language L, of entropy $H(L)$, there is always an e.d. ψ which is optimal in the sense that $\mu_\psi(L) = <\mu_\psi>(L) = H(L)/\ln_2 d$, relative to which there exists an absolutely-optimal effective encoder ψ_{-1}.

A class of decoders very important from a theoretical and practical point of

view is those of sequential-decoders which are such that if a word ξ of a language L can be factorized in subwords ξ_i (i=1,..,k) still belonging to L, then a code-word (or program) for ξ can be obtained by making the juxtaposition of the code-words of ξ_i's. This condition is very important in information theory since one is interested in transmitting sequences of words (or messages) of a given language.

Definition 2.3 .Let $\psi: Y^* \to X^*$ be an [effective]-decoder of a language L. It is called sequential if for all $k \in N$:

$$\psi(p_{i_1}),\ldots,\psi(p_{i_k}) \in L \text{ and } \psi(p_{i_1})\cdots\psi(p_{i_k}) \in L \Rightarrow \psi(p_{i_1}..p_{i_k}) = \psi(p_{i_1})\ldots\psi(p_{i_k})$$

From the definition it follows that if ψ is a sequential decoder of L and $\xi = \xi_{i_1}\ldots\xi_{i_k}$ with $\xi, \xi_{i_1}, \ldots \xi_{i_k} \in L$ then one has $K_\psi(\xi) \leq \sum_{i=1}^{k} K_\psi(\xi_i)$. It is possible to show that for any r.e.language L of entropy H(L) there is an effective sequential decoder ψ of L^* (L^* denotes the monoid generated by L) and therefore of L, which is optimal in the sense that $\mu_\psi(L) = <\mu_\psi>(L) = H(L)/\ln_2 d$.

3. Concluding remarks

In the setting of the communication-schema described in the introduction we have seen in the previous section some results on coding which are obtained by means of the Kolmogorov program-complexity theory. However we stress that the "efficiency" of such a coding does not depend only on the compression-coefficient of the words which one wishes to transmit, but also on the time of computation required to obtain them. In fact it can occur that one can keep "small" the amount of program but increasing the computation resources (time, space, etc) beyond any realistic limitation. Therefore also the "dynamic" aspects of the computation have to play a relevant role in this theory. Moreover we believe that the previous approach in which the receiver is schematized by an algorithm (or, more generally, by a formal system) can be a good frame to analyze higher levels of the communication as, for instance, how to transmit in order that the message affects the conduct of the receiver in the desidered way (pragmatical-level).

References

1 .Banerji,R.B.-Phrase structure languages, finite machines and channel capacity. Information and Control, 6, 153-162, 1963.

2 .Chaitin,G.J.-On the length of programs for computing finite binary sequences. J.ACM, 13, 547-569, 1966.

3 .Chomsky,N.and G.A.Miller,Finite state languages.Information and Control,1,91-112, 1958.

4 .De Luca,A.,Complexity and information theory.Proc.of summer school on"Coding and complexity"CISM,Udine,July 15-26,1974.

5 . " " .,Some information-theoretic aspects of the complexity theory,Proc.of the informal meeting on "Computational complexity,codes and formal languages".Laboratorio di Cibernetica,Naples,March 13-14,1975.

6 .De Luca A.and E.Fischetti,Outline of a new logical approach to information theory. In "New concepts and technologies in parallel information processing" (E.R.Caianiello ed.).Proc.of NATO summer school,Capri,June,17-30,1973. Noordhoff,Series E,n.9,1975.

7 .Kaminger,F.P.,The non computability of the channel capacity of context-sensitive languages".Information and Control,17,175-182,1970.

8 .Khinchin,A.I.,"Mathematical foundations of information theory".Dover Publ.1950

9 .Kolmogorov,A.N.,Three approaches to the quantitative definition of information. Problemy Pederachi Informatsii,1,3-11,1965.

10.Kuich,W.,On the entropy of context-free languages.Information and Control,16,173-200,1970.

11.Levin,L.A.and A.K.Zvonkin,The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms,Usphehi Mat.Nauk,25,85-127,1970

12.Mandelbrot,B.,On recurrent noise limiting coding,Proc.Symp.on Inf.Networks,Polytechn.Inst.of Brooklyn,205-221,1954.

13.Martin-Lof,P.,The definition of random sequences,Information and Control,6,602-619, 1966.

14.Salomaa,A."Formal languages".Academic Press.New York and London,1973.

15.Schnorr,C.P.,"Zufalligkeit und Wahrscheinlichkeit".Springer Verlag Lecture Notes in Mathematics,n.218,1970.

16.Shannon,C.E.,A mathematical theory of communication,Bell Syst.Tech.J.27,379-423, 1948.

17.Siromoney,R.,Channel capacity of equal matrix languages.Information and Control, 14,507-511,1969.

18.Solomonoff,R.J.,A formal theory of inductive inference.Part 1,Information and Control,7,1-22,1964.