Malaysian Journal of Science 21: 77-83 (2002)

Estimating the entropy of Chinese using the sliding-window entropy estimator

Choon-Peng Tan[1] and **Saw-Teng Yap**[2]

[1]Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]School of Arts and Science, Tunku Abdul Rahman College, Jalan Genting Kelang, 50932 Kuala Lumpur, Malaysia

ABSTRACT Three different text sources, namely a Chinese newspaper, the classical novel "Red Chamber Dream" and the modern prose "The Sahara" are selected for small-sample studies of the entropy of Chinese. We use the sliding window entropy estimator with the window size fixed at 1000 characters. By varying the number of window shifts up to 1000, we obtain entropy estimates of Chinese for the three different text sources. To improve the slow rate of convergence of the sliding window entropy estimator, we adopt the restricted sliding window estimator due to Kontoyiannis et al. Experimental indications are that modern Chinese has an entropy of less than 4.5 bits/character and that this entropy is less than that of classical Chinese.

ABSTRAK Tiga sumber teks berlainan, iaitu sebuah akhbar Cina , sebuah novel klasik "Red Chamber Dream" dan sebuah prosa moden "The Sahara" telah dipilih untuk kajian entropi Cina dengan sampel kecil. Kami menggunakan penganggar entropi tingkap bergerak dengan saiz tingkap ditetapkan pada 1000 aksara. Dengan mengubahkan bilangan pengalihan sehingga 1000, kami memperolehi anggaran entropi bahasa Cina untuk tiga sumber teks berlainan tersebut. Untuk memperbaiki kadar penumpuan penganggar entropi tingkap bergerak yang lambat, kami menggunakan penganggar tingkap bergerak terhad yang diperkenalkan oleh Kontoyiannis dan rakan sekerjanya. Hasil ujikaji menunjukkan bahawa bahasa Cina moden mempunyai entropi kurang daripada 4.5 bit/aksara dan entropi ini adalah kurang daripada entropi bahasa Cina klasik.

(small-sample, entropy of Chinese, sliding-window entropy estimator)

INTRODUCTION

Wyner and Ziv (1989) and Ornstein and Weiss (1993) have established an important result regarding the asymptotic behaviour of the phrase length in Lempel-Ziv coding (Cover and Thomas (1991), Ziv and Lempel (1977, 1978)). Before we state this result, we begin with some preliminaries.

Let the source $\{X_i\}$ be a stationary, ergodic process with values from a finite alphabet A. We let $\mathbf{X}_i^j$ denote the string $(X_i, X_{i+1}, \ldots, X_j)$ for $i \leq j$. For $n \geq 1$ and $i \geq 0$, we define $L_n(T^i\mathbf{X})$ to be the minimum length k such that the string $\mathbf{X}_i^{i+k-1}$ starting at position i does not appear as a continuous substring within the window $\mathbf{X}_{i-n}^{i-1}$ consisting of n symbols, where T denotes the shift operator. The following relationship is easy to establish:

$$L_n(T^i\mathrm{X}) = 1 + \max\left\{ \begin{array}{l} l: 0 \le l \le n,\ \mathrm{X}_i^{i+l-1} \\ = \mathrm{X}_{i-j}^{i-j+l-1} \ \textit{for some}\ l \le j \le n \end{array} \right\} \quad (1)$$

or, phrase length = 1+ maximum match-length. We note that in Lempel-Ziv coding, the shortest string $\mathbf{x}_i^{i+k-1}$ starting at position i that has not appeared as a substring of the window $\mathbf{x}_{i-n}^{i-1}$ is the next string or *phrase* to be encoded. Hence $L_n(T^i\mathbf{x})$ denotes the length of this phrase. Due to the relationship (1), we also call $L_n(T^i\mathbf{x})$ the

modified match-length. Associated with $L_n(T^i\mathbf{X})$, we define the maximum match-length $L_n^{(i)}(T^i\mathbf{X})$ as:

$$L_n^{(i)}(T^i\mathrm{X}) = \max\left\{\begin{array}{l} l:\ 0\le l\le n,\ \mathrm{X}_i^{i+l-1} \\ =\mathrm{X}_{i-j}^{i-j+l-1}\ \ for\ some\ l\le j\le n \end{array}\right\} \quad (2)$$

and hence by (1),

$$L_n(T^i\mathbf{X}) = 1 + L_n^{(i)}(T^i\mathbf{X}). \quad (3)$$

By stationarity, we can shift the window $\mathbf{X}_{i-n}^{i-1}$ to $\mathbf{X}_k^{k+n-1}$, with $i = k + n$, an alternative expression for $L_n^{(i)}(T^i\mathbf{X})$ is:

$$L_n^{(k+n)}(T^{k+n}\mathrm{X}) = \max\left\{\begin{array}{l} l:\ 0\le l\le n,\ \mathrm{X}_{k+n}^{k+n+l-1} \\ =\mathrm{X}_{k+n-j}^{k+n-j+l-1}\ \ for\ some\ l\le j\le n \end{array}\right\}$$

$$= \max\left\{\begin{array}{l} l:\ 0\le l\le n,\ \mathrm{X}_{k+n}^{k+n+l-1} = \mathrm{X}_s^{s+l-1} \\ for\ some\ \ k\le s\le k+n-l \end{array}\right\} = L_n^{(k)}(T^k\mathbf{Y})$$

where $\mathbf{Y} = T^n\mathbf{X}$.

For a window of fixed size n and at $\mathbf{X}_k^{k+n-1}$ after $k-1$ shifts of this window, we shall use the notation $L^{(k)}$ to denote the *maximum match-length* after $k-1$ shifts of the window, namely:

$$L^{(k)} = \max\left\{\begin{array}{l} l:\ 0\le l\le n,\ \mathrm{X}_{k+n}^{k+n+l-1} = \mathrm{X}_s^{s+l-1} \\ for\ some\ \ k\le s\le k+n-l \end{array}\right\}, \quad (4)$$

as a consequence of the ergodic property. We are now ready to state the Wyner-Ziv-Ornstein-Weiss result as follows:

Wyner-Ziv-Ornstein-Weiss Theorem

Let $\{X_i\}$ be a stationary ergodic process. For any starting time i, $\frac{L_n(T^i\mathbf{X})}{\log n} \to \frac{1}{H}$ a.s. as $n \to \infty$, where H is the entropy rate of $\{X_i\}$ and $L_n(T^i\mathbf{X})$ is the phrase length.

Corollary. Let $\{X_i\}$ be a stationary ergodic process. For any starting time i, $\frac{L_n^{(i)}(T^i\mathbf{X})}{\log n} \to \frac{1}{H}$ a.s. as $n \to \infty$, where $L_n^{(i)}(T^i\mathbf{X})$ is the maximum match-length and H is the entropy rate of $\{X_i\}$.

Remarks.

(i) Wyner and Ziv (1989) basically established convergence in probability. The complete proof of almost sure convergence was given by Ornstein and Weiss (1993).

(ii) The result also holds for

$$\hat{L}_n(T^i\mathrm{X}) = 1 + \max\left\{\begin{array}{l} l:\ \mathrm{X}_i^{i+l-1} = \mathrm{X}_{i-j}^{i-j+l-1} \\ for\ some\ \ 1\le j\le n \end{array}\right\} \quad (5)$$

namely:

$$\frac{\hat{L}_n(T^i\mathbf{X})}{\log n} \to \frac{1}{H}\ \text{ a.s.}\ \text{ as } n\to\infty. \quad (6)$$

Similarly, for

$$\hat{L}_n^{(i)}(T^i\mathrm{X}) = \max\left\{\begin{array}{l} l:\ \mathrm{X}_i^{i+l-1} = \mathrm{X}_{i-j}^{i-j+l-1} \\ for\ some\ \ 1\le j\le n \end{array}\right\} \quad (7)$$

we have the corresponding result:

$$\frac{\hat{L}_n^{(i)}(T^i\mathbf{X})}{\log n} \to \frac{1}{H}\ \text{ a.s.}\ \text{ as } n\to\infty, \quad (8)$$

where $\mathbf{x}_i^{i+l-1}$ is allowed to match with a string beginning in the window $\mathbf{x}_{i-n}^{i-1}$, i.e. $\mathbf{x}_i^{i+l-1}$ need not be a substring of the window.

Consider the following estimator introduced by Farach et al. (1995)

$$F_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\frac{L^{(i)}}{\log n} \quad (9)$$

where the maximum match-length $L^{(i)}$ is defined by (4). By the Wyner-Ziv-Ornstein-Weiss Theorem, $\frac{L^{(i)}}{\log n} \to \frac{1}{H}$ a.s. as $n \to \infty$, where H is the entropy rate of the underlying stationary, ergodic process.

For a fixed n, the Cesaro average

$$F_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\frac{L^{(i)}}{\log n} \to E\left(\frac{L^{(i)}}{\log n}\right)\ \text{ a.s. as } k\to\infty$$

by the ergodic property. Wyner (1993, 1997) has shown that, for a stationary ergodic finite-order Markov process, and for all i,

$$E(L^{(i)}) = \frac{\log n}{H} + O(1) \tag{10}$$

as $n \to \infty$. Equivalently,

$$E\left(\frac{L^{(i)}}{\log n}\right) = \frac{1}{H} + o(1) \tag{11}$$

as $n \to \infty$. Hence if $k \to \infty$ and then n $\to \infty$,

$$F_{k,n} \to \frac{1}{H} \text{ a.s.} \tag{12}$$

Defining

$$\hat{H}_{k,n} = \frac{1}{F_{k,n}} = \frac{k \log n}{\sum_{i=1}^{k} L^{(i)}}, \tag{13}$$

we have

$$\hat{H}_{k,n} \to H \text{ a.s. as } k \to \infty \text{ and } n \to \infty. \tag{14}$$

Thus $\hat{H}_{k,n}$ is the *sliding-window estimator* estimating the entropy rate H (Farach et al. (1995), Wyner (1995)). Kontoyiannis et al. (1998) have shown that (11) is true for any stationary ergodic process that satisfies the following condition:

$$E\left\{\sup_n \frac{L_n(T^i \mathbf{X})}{\log n}\right\} < \infty \tag{15}$$

for all i. The *Doeblin Condition* (DC) below is a sufficient condition for (15) to be satisfied. *Doeblin Condition* (DC): There exists an integer $r \geq 1$ and a real number $\beta \in (0,1)$ such that, for all $x_0 \in A$ (the alphabet), $P\{X_0 = x_0 \mid \mathbf{X}_{-\infty}^{-r}\} \leq \beta$, with probability one.

Hence, for any stationary ergodic process satisfying the Doeblin Condition, (11) is true and $\hat{H}_{k,n} \to H$ a.s. as $k \to \infty$ and $n \to \infty$ where the sliding-window estimator $\hat{H}_{k,n}$ is defined by (13). We note that the Doeblin condition is satisfied for independent, identically distributed (i.i.d.) processes, stationary ergodic Markov chains of finite order and certain non-Markov processes. This condition is satisfied by natural languages and many practical sources of data.

2. Sliding-Window Entropy Estimation of Chinese

Assume that the source $\{X_i\}$ is a stationary and ergodic process satisfying the Doeblin condition. The entropy rate H of the process is defined as:

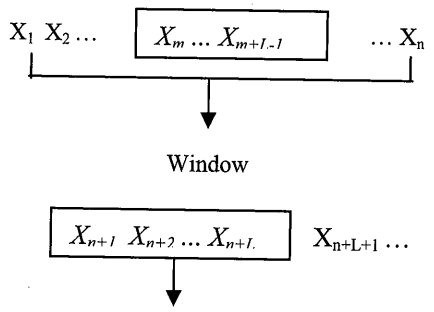
$$H = \lim_{K \to \infty} \frac{H(X_1,...,X_K)}{K} \tag{16}$$

where $H(X_1,...,X_K)$ is the joint entropy of $X_1,...,X_K$. The sliding-window estimator of H is:

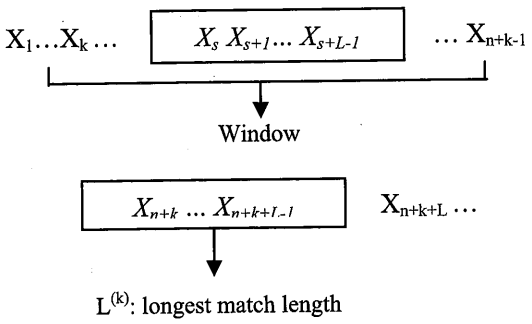
$$\hat{H} \stackrel{\Delta}{=} \frac{K \log_2 n}{\sum_{k=1}^{K} L^{(k)}} \tag{17}$$

where n is the size of a given window, K is a large integer and $L^{(k)}$ is the longest match length corresponding to k as defined below. Observe $\{X_1, X_2,...\}$ and let $\mathbf{X}_1^n$ be a given window. Then for k =1, $L^{(1)}$ is the largest integer L such that $\mathbf{X}_{n+1}^{n+L} = \mathbf{X}_m^{m+L-1}$, for some m $\in [1, 1+n-L]$.

See diagram below:



Then shift the window 1 position so that the new window is $\mathbf{X}_2^{n+1}$. Continue shifting until the window is $\mathbf{X}_k^{n+k-1}$ (after $k-1$ shifts). Define $L^{(k)}$ as the largest integer L such that $\mathbf{X}_{n+k}^{n+k+L-1} = \mathbf{X}_s^{s+L-1}$ for some $s \in [k, k+n-L]$, where k = 1,2,3,.... See diagram below:



In the sequel, we assume that printed Chinese is a stationary and ergodic process satisfying the Doeblin condition. The sliding-window estimator $\hat{H}$ given by (17) is used to estimate the entropy rate H of Chinese. We select the following 3 text sources for our experimental study.

Text 1. This is an excerpt from a classical Chinese novel '*Red Chamber Dream*'. This well-known novel is believed to be written by Cao Xue Qin with the assistance of a group of intellectuals during the Ching Dynasty. It consists of 120 chapters. The first 2000 characters of the book form the experiment text.

Text 2. This is a modern Chinese prose '*The Sahara*' written by San-Mao in 1976. The first 2000 characters of the book form the experimental text.

Text 3. This is a collection of murder cases reported in the Chinese newspaper, Sin Chew Jit Poh for the month of May, 1998. The news reports were collected and arranged according to the date and page sequence which appeared in the newspaper. A total of 2000 characters were used as the text source.

Note that in all the 3 texts, all punctuation marks and spaces, including non-Chinese characters that have appeared are removed. In other words, these are not regarded as source symbols. The experimental results are shown in Tables 1 and 2 and Figure 1.

Table 1. Entropy estimate of Chinese for the 3 different text sources based on different values of K, the number of window shifts, and $\sum_{k=1}^{K} L^{(k)}$ (the total match length) where the window size is fixed at $n = 1000$ characters.

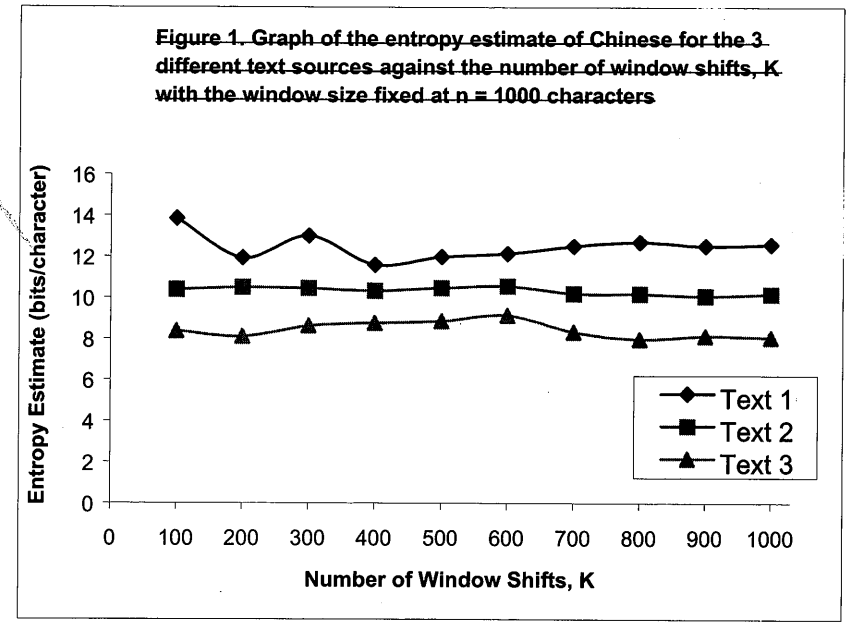
	Text 1		Text 2		Text 3	
Number of Window Shifts K	**Total Match Length** $\sum_{k=1}^{K} L^{(k)}$	**Entropy Estimate (bits/ character)** $\hat{H}$	**Total Match Length** $\sum_{k=1}^{K} L^{(k)}$	**Entropy Estimate (bits/ character)** $\hat{H}$	**Total Match Length** $\sum_{k=1}^{K} L^{(k)}$	**Entropy Estimate (bits/ character)** $\hat{H}$
100	72	13.84	96	10.38	119	8.37
200	167	11.94	190	10.49	246	8.10
300	230	13.00	286	10.45	347	8.62
400	344	11.59	386	10.33	455	8.76
500	416	11.98	476	10.47	563	8.85
600	493	12.13	566	10.56	655	9.13
700	558	12.50	684	10.20	838	8.32
800	628	12.70	783	10.18	999	7.98
900	717	12.51	890	10.08	1102	8.14
1000	793	12.57	980	10.17	1237	8.06

Table 2. Entropy estimates of Chinese based on 3 different text sources using the sliding-window method where the window size is fixed at $n = 1000$ characters.

Text	Number of window shifts K	Total Match Length $\sum_{k=1}^{K} L^{(k)}$	Entropy Estimate (bits/character) $\hat{H}$
1	1000	793	12.57
2	1000	980	10.17
3	1000	1237	8.06

Table 3. Entropy estimate of Chinese based on 3 different text sources using the restricted sliding-window estimator where the window size is fixed at $n = 1000$ characters.

Text	Total Modified Match Length $\sum_{i=1}^{n} L_n(T^i \mathbf{x})$	Entropy Estimate (bits/character) $\hat{H}_n$
1	1793	5.56
2	1980	5.03
3	2237	4.45



Interpretation of the experimental results

From Table 1, we observe that for Text 1 with a window size of 1000 and 1000 shifts of the window, the total match length obtained is 793 and the entropy estimate is 12.57 bits/character. From the same table, we find that for Text 2 with the same window size and the same number of window shifts, the total match length obtained is 980 and the entropy estimate is 10.17 bits/character. The entropy estimated for Text 2 is lower than the entropy estimated for Text 1. We observe that for Text 3 with 1000 window shifts, the total match length is 1237 and the entropy estimate is 8.06 bits/character. This

entropy estimate is the lowest among the three texts.

From Table 2, we observe that the entropy estimate based on a classical Chinese text source (i.e. text 1), namely 12.57 bits/character, is higher than that of modern Chinese text sources (i.e. texts 2 and 3 namely 10.17 bits and 8.06 bits respectively) when $n = K = 1000$. In general, Table 1 and Figure 1 show that for other fixed values of n and K, the entropy estimate of classical Chinese is higher than that of modern Chinese. Classical Chinese text contains more Chinese characters which are rarely used. In contrast, modern Chinese text like newspaper sources contain many characters which are in daily use.

The convergence rate of the sliding-window entropy estimate is quite slow (see Figure 1). To obtain a better entropy estimate, we need to increase the window size n and the number of window shifts K. A better estimate of the entropy of Chinese, namely 4.1 bits/character was obtained by Victor Wei (1986) who adopted the gambling approach of Cover and King (1978).

We can improve the entropy estimate by using the restricted sliding-window estimator proposed by Kontoyiannis et al. (1998) recently.

3. The Restricted Sliding-Window Estimator Based On Phrase Lengths

Kontoyiannis et al. (1998) proposed a restriction of the sliding-window entropy estimation method to estimate the entropy of English. Based on a sample of about 75,000 words from the novel *Mansfield Park* by Jane Austen, an estimate of 1.777 bits/ character was obtained. They also showed that their estimators are very efficient for small sample sizes.

Let $\{X_i\}$ be a stationary and ergodic process. Then the entropy rate H of the process is estimated by

$$\hat{H}_n = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{L_n(T^i \mathbf{x})}{\log_2 n} \right]^{-1} \qquad (18)$$

assuming that the process satisfies the Doeblin condition, where n (a large integer) is the size of a given window and $L_n(T^i\mathbf{X})$ the phrase length is defined below. Wyner (1993, 1997) proved the convergence of the sliding-window estimator (17) to the entropy rate H of the process if the process is finite order Markov. The *restricted sliding-window* estimator (18) with $K = n$ is proposed by Kontoyiannis et al. (1998). They proved the almost-sure convergence of the restricted sliding-window estimator to the entropy rate H if the Doeblin Condition is satisfied. For $n > 1$ and $i = 0$, $L_n(\mathbf{X})$ is defined as the minimum length k such that the string $\mathbf{X}_0^{k-1}$ that starts at time 0 does not appear as a continuous substring of the past string $\mathbf{X}_{-n}^{-1}$.
In fact,

$$L_n(\mathrm{X}) = 1 + \max \left\{ \begin{array}{c} l:\ 0 \le l \le n,\ \mathrm{X}_0^{l-1} = \mathrm{X}_{-j}^{-j+l-1} \\ \textit{for some } l \le j \le n \end{array} \right\}. \qquad (19)$$

For $n > 1$ and $i \square 0$, $L_n(T^i\mathbf{X})$ is defined as the minimum length k such that the string $\mathbf{X}_i^{i+k-1}$ starting at position i does not appear as a continuous substring of the previous string $\mathbf{X}_{i-n}^{i-1}$ consisting of n symbols (i.e. $\mathbf{X}_{i-n}^{i-1}$ is the corresponding window). It is easier to use the following relationship to calculate $L_n(T^i\mathbf{X})$:

$$L_n(T^i\mathrm{X}) = 1 + \max \left\{ \begin{array}{c} l:\ 0 \le l \le n,\ \mathrm{X}_i^{i+l-1} = \mathrm{X}_{i-j}^{i-j+l-1} \\ \textit{for some } l \le j \le n \end{array} \right\}. \qquad (20)$$

Due to (20), the restricted sliding-window estimator (18) converges to the entropy rate H faster than the sliding-window estimator (17). Using phrase lengths $L_n(T^i\mathbf{X})$ instead of match lengths $L^{(k)}$ is more efficient in entropy estimation.

Using the restricted sliding-window estimator, we obtain the following results in Table 3 based on the previous experimental data.

These results show considerable improvement in the entropy estimates. From Table 3, we observe that the entropy estimate based on a classical Chinese text (i.e. Text 1), namely 5.56 bits/character, is higher than that of a modern Chinese text source (i.e. Text 3), namely 4.45 bits/character, when $n = K = 1000$. The entropy estimate of 4.45 bits/character for Text 3 is closer to the estimate obtained by Victor K. Wei, namely 4.1 bits/character. Our text source 3 is based on news reports written in modern Chinese

which is closer to the text source of modern science fiction studied by Wei.

We find that the entropy estimates obtained by the restricted estimator are lower than that obtained by the sliding-window estimator for all the three texts. The entropy estimate of Text 1 shows an improvement of 55.77% with the estimate going down to 5.56 bits/character. Text 2 and Text 3 show improvements of 50.54 % and 44.79% respectively. Experimental indications are that modern Chinese has an entropy of less than 4.45 bits/character and that this entropy is less than that of classical Chinese. These small-sample estimates compare favourably with the estimate of 4.1 bits/character obtained by Victor Wei (1986) using the entropy estimation scheme of Cover and King (1978).

REFERENCES

1. Cover, T. M. and King, R. (1978). "A Convergent Gambling Estimate of the Entropy of English," *IEEE Trans. Inform. Theory* **24**: 413-421.
2. Cover, T. M. and Thomas, J. (1991). *Elements of Information Theory*, New York: Wiley.
3. Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A. J. and Ziv, J. (1995). "On the Entropy of DNA: Algorithms and Measurements Based on Memory and Rapid Convergence," in *Proc. ACM-SIAM Symp. Discrete Algorithms. Philadelphia,* PA: Soc. Industr. Appl. Math. p. 48-57
4. Kontoyiannis, Algoet, P. H., Suhov, Yu. M. and Wyner, A. J. (1998). "Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text," *IEEE Trans. Inform. Theory* **44**: 1319-1327
5. Ornstein, D. S. and Weiss, B. (1993). "Entropy and Data Compression Schemes," *IEEE Trans. Inform. Theory* **39**: 78-83.
6. Wei, Victor K. (1986). "The Information Entropy of the Chinese Language," *Conference Record of the International Symposium on Information Theory,* Ann Arbor, Michigan, U.S.A p. 96
7. Wyner, A.D. (1995). "Typical Sequences and All That: Entropy, Pattern Matching, and Data Compression," *IEEE Trans. Inform. Theory Society Newsletter*, p. 8-15.
8. Wyner, A.D. and Ziv, J. (1989). "Some Asymptotic Properties of Entropy of a Stationary Ergodic Data Source with Applications to Data Compression," *IEEE Trans. Inform. Theory* **35**: 1250-1258.
9. Wyner, A.J. (1993). "String matching theorems and applications to data compression and statistics," Ph.D. dissertation, Stanford University, U.S.A.
10. Wyner, A.J. (1997). "The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm," *IEEE Trans. Inform. Theory* **43**: 1452-1464.
11. Ziv, J. and Lempel, A. (1977). "A Universal Algorithm for Sequential Data Compression," *IEEE Trans. Inform. Theory* **23**: 337-343.
12. Ziv, J. and Lempel, A. (1978). "Compression of Individual Sequences by Variable Rate Coding," *IEEE Trans. Inform. Theory* **24**: 530-536.