Lower Bounding the Optimal $LZ78$-Parsing$^1$

Marcelo S. Pinho$^2$ & Weiler A. Finamore$^3$
mpinho@feg.unesp.br weiler@cetuc.puc-rio.br
$^2$ Universidade Estadual Paulista, Guaratinguetá - S.P., Brazil
$^3$ Pontificia Universidade Católica, Rio de Janeiro - R.J., Brazil

The $LZ78$ algorithm, a lossless data compressor based on string matching, was proposed in 1978 by Ziv and Lempel as an elegant solution to the problem of universal source coding. Although it is asymptotically optimal for any ergodic source, that is, its compression rate converges to the source entropy, the asymptotic behavior of this compression rate is not optimal for many classes of sources, such as memoryless, markovian and finite state machine (fsm) sources [1]. Therefore, the performance of the $LZ78$ can be improved.

Given an input sequence $x_1^n$ (a realization of source output), the $LZ78$ parses this sequence into phrases, based on an adaptative dictionary, and encodes the phrase sequentially. It is well known that the $LZ78$ parsing procedure is not optimal. However, in [2] it was shown that the optimal parsing procedure is NP-complete and its performance is unknown. Therefore the gain which can be obtained by using a better parsing procedure in the $LZ78$ is also unknown. In this work this problem is studied.

The number of phrases produced by the $LZ78$ parsing procedure is related to its compression rate. In fact, the problem of optimizing the compression rate is equivalent to the problem of finding a parsing procedure with minimum number of phrases. Our first result shows that the (unknown) number of phrases produced by the optimal $LZ78$-parsing has the same asymptotic behavior as the number of phrases produced by the $LZ78$-parsing. This result is stated in the following theorem.

**Theorem:** Let $x_1^n$ be a sequence drawn from an ergodic source. The optimal $LZ78$-parsing of $x_1^n$ produces a number $m_{opt}$ of phrases which is related to the number $m_{LZ78}$ of phrases produced by $LZ78$-parsing by $\lim_{n \to \infty} \frac{m_{opt}}{m_{LZ78}} = 1$, as.

Although the optimal $LZ78$-parsing cannot be computed easily, we developed a simple sequential algorithm which can compute a lower bound to the number of phrases produced by the optimal parsing procedure. This algorithm was tested and the results are compared to the number of phrases produced by two versions of the $LZ78$: (i) the well known $LZW$ and (ii) the algorithm $mm-LZ78$, introduced in [3]. The results have shown that the lower bound is close to the number of phrases of those versions (10% better in the Canterbury Corpus).

References


$^1$This work was supported by CNPq - Brazil