Symbol Ranking text compressors

Peter Fenwick,
Department of Computer Science, The University of Auckland,
Private Bag 92019, Auckland, New Zealand.
peter-f@cs.auckland.ac.nz

In 1951 Shannon estimated the entropy of English text by giving human subjects a sample of text and asking them to guess the next letters. He found, in one example, that 79% of the attempts were correct at the first try, 8% needed two attempts and 3% needed 3 attempts. By regarding the number of attempts as an information source he could estimate the language entropy.

Shannon also stated that an “identical twin” to the original predictor could recover the original text and these ideas are developed here to provide a new taxonomy of text compressors. In all cases these compressors recode the input into “rankings” of “most probable symbol”, “next most probable symbol”, and so on. The rankings have a very skew distribution (low entropy) and are processed by a conventional statistical compressor. Several “symbol ranking” compressors have appeared in the literature, though seldom with that name or even reference to Shannon’s work. More recently, some unusual compressors have appeared, some with very high performance, and many of these can be classified as symbol ranking.

Compressors of this type have been described by

2. Lelewer and Hirschberg 1991 Data Compression Conference, DCC-91
3. Howard and Vitter 1993 Data Compression Conference, DCC-93
5. Buynovski 1994 "Monitor", Moscow, No 8 (ACB compressor)

The last reference is apparently the first compressor to be specifically designed as “symbol ranking”. While symbol ranking has a general resemblance to PPM, it has the unique feature that the context order is never released to the final coder and that there are no escapes between orders.

While most of the compressors in the list use variable-order contexts, the author has developed one which uses constant-order contexts and is based on a set-associative cache with LRU update. Cache lines are addressed by hashing the last 3 symbols to give an order-3 context. The sets hold the most-recent symbols within the context, managed according to a conventional LRU mechanism; the LRU position is emitted as the symbol rank. Less-frequent symbols are handled as positions in a pseudo-MTF list.

A software implementation has run at about 1 Mbyte/s with an average compression of 3.6 bits/byte on the Calgary Corpus.