Text Compression as a Test for Artificial Intelligence

Matthew V. Mahoney
415 Rutgers Ave., Melbourne FL 32901, 407-724-1582
matmahoney@aol.com, http://www.he.net/~mmahoney/
Florida Institute of Technology, Computer Science Dept.

It is shown that optimal text compression is a harder problem than artificial intelligence as defined by Turing's (1950) imitation game; thus compression ratio on a standard benchmark corpus could be used as an objective and quantitative alternative test for AI (Mahoney, 1999). Specifically, let L, M, and J be the probability distributions of responses chosen by a human, machine, and human judge respectively to the judge's questions in the imitation game. The goal of AI is $M = L$, the machine is indistinguishable from human. But the machine wins (the judge guesses that it is human) when $H_J(M) < H_J(L)$, where $H_Q(P) \equiv -\Sigma_x P(x) \log Q(x)$ is the cross entropy of Q with respect to P. This happens when J is a poor estimate of L, meaning that the interrogator fails to anticipate the human's responses, but even in the worst case when $J = L$, the machine can still win with a suboptimal solution ($M \neq L$) by deterministically favoring the most likely responses over the true distribution. In contrast, optimal compression of a probabilistic language L with unknown distribution (such as English) using an estimated distribution M (an encoding of length $-\log_2 M(x)$ bits for each string x) is $M = L$, by the discrete channel capacity theorem (Shannon, 1949).

Answering questions in the Turing test (*What are roses?*) seems to require the same type of real-world knowledge that people use in predicting characters in a stream of natural language text (*Roses are ___?*), or equivalently, estimating $L(x)$ for compression. Shannon (1951), and Cover and King (1978) established an upper bound of 1.3 bits per character (bpc) for the entropy (information content) of English narrative in a 27-character alphabet (A-Z and space) using human prediction tests.

No compression program has achieved this. Eight programs, including those top-rated by Gilchrist (1998) and Bell (1998) were used to compress English narrative, *Alice in Wonderland* (Carroll, 1865) and *Far from the Madding Crowd* by Thomas Hardy (*book1* from the Calgary corpus (1993)), after reducing both to 27 characters. The best compression was achieved by *rkive* (Taylor, 1998): 1.86 bpc on *alice* and 1.94 on *book1*. Others tested (from worst to best) were *compress* (1990), *pkzip* (1993), *gzip* (Gailly, 1993), *ha* (Hirvola, 1993), *szip* (Schindler, 1998), *ppmz* (Bloom, 1998), and *boa* (Sutton, 1998). All program options were set for maximum compression.

Better compressors “learn”, using prior input to improve compression on subsequent input. *szip* was the best learner, compressing *book1* to about 95% of the size of the two halves compressed separately. Fig. 1 shows the correlation between compression and learning. Similar results were obtained for *alice*.

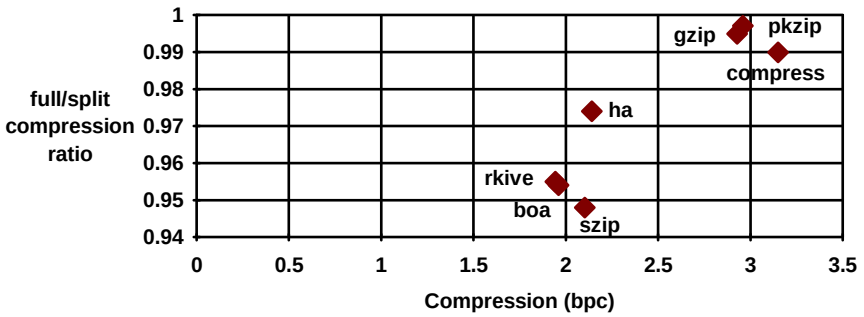

Fig. 1. Full, and ratio of full to split compression for *book1* using a 27 character alphabet.

It was also found that better compressors make greater use of the syntactic and semantic constraints of English. Lexical, syntactic, and semantic constraints were selectively broken by swapping pairs of letters within words, pairs of words, or pairs of phrases respectively. Results for the original text of *book1* are shown in Fig. 2, with similar results for *alice*. The swapping transforms are reversible and do not change file size or information content.

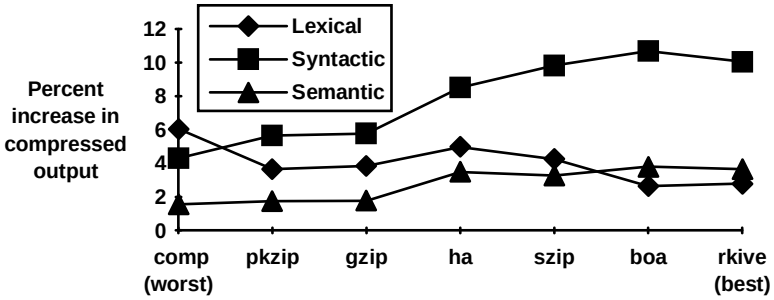

Fig. 2. Percent increase in compressed output for *book1* as compression improves when lexical, syntactic, or semantic constraints are selectively broken.

Acknowledgments. I would like to thank Dr. Phil Chan for guidance in developing this paper.

References

Bell, T., 1998. Canterbury Corpus, http://corpus.canterbury.ac.nz/

Bloom, C., 1998. Solving the Problems of Context Modeling. http://www.cco.caltech.edu/~bloom/papers/ppmz.zip

Calgary Corpus 1993. http://www.kiarchive.ru/pub/msdos/compress/calgarycorpus.zip

Carroll, L., 1865. *Alice in Wonderland*. Gutenberg Press, ftp://sunsite.unc.edu/pub/docs/books/gutenberg/etext97/alice30h.zip

compress 4.3d for MSDOS, 1990. ftp://ctan.tug.org/tex-archive/tools/compress/msdos.zip

Cover, T. M., and King, R. C., 1978. A Convergent Gambling Estimate of the Entropy of English. IEEE Transactions on Information Theory 24:413-421.

Gailly, J. 1993. gzip 1.2.4, http://www.kiarchive.ru/pub/msdos/compress/gzip124.exe

Gilchrist, J. 1998. Archive Comparison Test, http://www.geocities.com/SiliconValley/Park/4264/act-mcal.html

Hirvola, H., 1993. ha 0.98, http://www.webwaves.com/arcers/msdos/ha098.zip

Mahoney, M., 1999. Text Compression as a Test for Artificial Intelligence, submitted for publication, http://www.he.net/~mmahoney/paper4.ps.Z

PKZIP 1993, version 2.04e, PKWARE Inc.

Rich, E, and Knight, K., 1991. *Artificial Intelligence*, 2nd Ed., New York: McGraw-Hill.

Schindler, M., 1998. szip homepage, http://www.compressconsult.com/szip/

Shannon, C., and Weaver W., 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Shannon, C. 1951. Prediction and Entropy in Printed English. Bell Sys. Tech. J 3:50-64.

Sutton, I., 1998. boa 0.58 beta, http://webhome.idirect.com/~isutton/

Taylor, M., 1998. RKIVE v1.91 beta 1, http://www.geocities.com/SiliconValley/Peaks/9463/rkive.html