Empirical Evidence for Hilberg's Conjecture in Single-Author Texts

Łukasz Dębowski
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
ldebowsk@ipipan.waw.pl

Abstract

Hilberg's conjecture is a statement that the mutual information between two adjacent blocks of text in natural language scales as n^β, where n is the block length. Previously, this hypothesis has been linked to Herdan's law on the levels of word frequency and of text semantics. Thus it is worth a direct empirical test. In the present paper, Hilberg's conjecture is tested for a selection of English prose using the Lempel-Ziv algorithm. An upper bound for the exponent β is found to be 0.949.

Keywords: single-author texts, universal coding

1 Introduction

Texts typically produced by humans diverge from both pure randomness and simple determinism. If we investigate predictability of such texts borrowing tools from information theory, we should observe some particular behavior of their optimal compression rate. Namely, the compression rate as a function of the text length should neither tend very fast to zero (the case of determinism) nor tend very fast to a constant greater than zero (the case of pure randomness). Concurring with this intuition, German telecommunications engineer Wolfgang Hilberg [7] supposed that the optimal compression rate of a text in natural language scales as $n^{-1+\beta}$, where n is the length of the text and β is close to 0.5. Hilberg's conjecture was motivated largely rationally but was partly based on an extrapolation of Shannon's seminal experimental data [10], which contained the estimates of conditional entropy for blocks of $n \leq 100$ characters.

As can be easily shown, Hilberg's conjecture implies that mutual information between two adjacent text blocks of length n is proportional to n^β. Using more involved mathematical modeling, the latter property can be linked with the distribution of words appearing in texts and the distribution of facts described by texts. First, Dębowski [4] has proved a theorem by which the power-law growth of mutual information implies that the number of distinct set phrases (words) in a text of length n roughly exceeds n^β divided by a logarithmic term, cf. Dębowski [5]. The claim of this theorem is actually observed and known as Herdan's law [6]. Second, Dębowski [4] has proved a proposition which says that the power-law growth of mutual information is obeyed if a text of length n describes more than n^β independent facts in a repetitive fashion. Hence

Hilberg's conjecture may be linked to power-laws on the levels of word frequency and of text semantics.

In view of these mathematical results, Hilberg's conjecture deserves experimental validation. Whereas it seems dubious that the optimal compression rate or the conditional entropy tends to zero for (con)text lengths tending to infinity, it is plausible that the mutual information between large adjacent text blocks grows according to a power law. Ever since Shannon the entropy of natural language has been the object of often scientific investigation but the exact scaling of the compression rate is a little investigated issue. Therefore, we decided to devote the present paper to the specific topic of verifying Hilberg's conjecture. The findings of this paper have a preliminary character.

Reviewing earlier research, we first mention Cover and King [2], who found an estimate of the asymptotic conditional entropy of English texts as 1.25 bpc (bits per character). This estimate was obtained using human subjects who were instructed to gamble on consecutive letters of the text and an estimate of entropy was computed from the accumulated capital. Modern compression algorithms compare with these estimates favorably. PPM (prediction by partial matching), being one of the best-performing compression algorithms, achieves the compression rate of 1.46 bpc for selected English texts [11]. Similar studies have been done for languages other than English, cf. e.g. Behr et al. [1], and for other compression algorithms, cf. Mahoney [9].

A graph depicting how PPM's compression rate depends on the amount of training text is also given by Teahan and Cleary [11]. We are looking for a bit different graphs, namely, how the compression rate and the block mutual information depend on the amount of compressed text. For this reason we have decided to perform an independent compression experiment. In any such experiment there are two variables to be fixed. The first one is the compression algorithm, the second is the selection of texts.

For simplicity, we evaluate the compression rate and the mutual information using the Lempel-Ziv code [12], which is the simplest of universal codes. Universal codes are compression algorithms which asymptotically get the optimal compression rate for stationary sources. It can be shown that the estimates of mutual information given by universal codes are greater than the true mutual information. Moreover, the difference between the estimate of mutual information and the true mutual information is the smaller, the better the compression rate is. Hence we may use a universal code to upper bound the true mutual information.

Another important issue is what is the range of texts for which Hilberg's conjecture can be reasonably verified. One can consider either single (i.e., non-concatenated) texts produced by single authors or concatenations of such texts (i.e., corpora). We have decided to consider first only single-author texts since the compression rate for concatenated texts may depend on the specific choice of the text collection. Thus we consider a selection of single-author texts in English downloaded from the project Gutenberg.

In a nutshell, the findings of this paper can be thus described. In the range of text lengths $n \in (10^3, 10^7)$ characters, we observe a power-law relationship for both the compression rate and the mutual information computed for the Lempel-Ziv code. The fitted exponent for the compression rate is close to $\beta \approx 0.949$. This observation does not exclude Hilberg's conjecture with a very high exponent β. However, if we used a better universal code then we might obtain

a tighter bound. For this reason it is advisable to repeat our experiment using better codes than the Lempel-Ziv, such as the PPM code.

The subsequent organization of the paper is as follows. In Section 2, we introduce some necessary concepts from information theory. In Section 3, we expose Hilberg's conjecture. In Section 4, we reformulate this conjecture using mutual information. In Section 5, we discuss the experiment. The paper is concluded in Section 6.

2 A bit of information theory

We first give a brief primer on information theory, cf. Cover and Thomas [3]. The fundamental concept of information theory is the entropy of a random variable. For a random variable $X_1^n = (X_1, X_2, ..., X_n)$, where X_i are consecutive characters of a random text, the entropy is defined as

$$H(X_1^n) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n). \tag{1}$$

If we have a uniquely decodable code C for variable X_1^n, then the expectation of its length $|C(X_1^n)|$ cannot be smaller than the entropy, i.e.,

$$\sum_{x_1^n} P(X_1^n = x_1^n) \, |C(x_1^n)| \geq H(X_1^n). \tag{2}$$

It can be shown that there exist a uniquely decodable code C with lengths $|C(x_1^n)| = \lceil -\log P(X_1^n = x_1^n) \rceil$, called the Shannon-Fano code. For this code we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n) \, |C(x_1^n)| \leq H(X_1^n) + 1. \tag{3}$$

The length of the Shannon-Fano code $\lceil -\log P(X_1^n = x_1^n) \rceil$ could be considered the information content of an individual text x_1^n.

However, we cannot evaluate the Shannon-Fano code if the proper probability distribution P is not specified or does not exist. As noticed by Kolmogorov [8], this may be well the case of natural language. In such a case, Kolmogorov proposed to define the information content of an individual text x_1^n as the length of the shortest program for a simple universal computer (a Turing machine) that makes the computer produce x_1^n on its output. This quantity is called Kolmogorov complexity $K(x_1^n)$. For any computable code C there exists a constant c such that

$$K(x_1^n) \leq |C(x_1^n)| + c. \tag{4}$$

Since Kolmogorov complexity is itself a length of a computable code, we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n) K(x_1^n) \geq H(X_1^n) \tag{5}$$

for a random variable X_1^n on a definite probability space. In case of a computable probability distribution the Shannon-Fano code is also computable so, from (3)

and (4), we obtain

$$\sum_{x_1^n} P(X_1^n = x_1^n) K(x_1^n) \leq H(X_1^n) + c + 1. \tag{6}$$

Hence the expectation of Kolmogorov complexity for computable distributions is close to entropy. In contrast, the difference between Kolmogorov complexity and entropy can be arbitrarily large for noncomputable distributions.

The problem about Kolmogorov complexity is, however, that it is not computable. Therefore we will rather take a middle path to measuring information content of individual texts, which is universal coding. A universal code is a uniquely decodable computable code C which for any stationary stochastic process $(X_1, X_2, ...)$ achieves the optimal compression rate

$$\lim_{n\to\infty} \frac{1}{n} \sum_{x_1^n} P(X_1^n = x_1^n) \, |C(x_1^n)| = h, \tag{7}$$

where the asymptotic entropy rate is

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1^n). \tag{8}$$

Some example of a universal code is the Lempel-Ziv code [12]. Subsequently, we will measure the information content of an individual text as the length of this code.

3 Flavors of Hilberg's conjecture

We are now in a position to introduce Hilberg's conjecture. The original form of this hypothesis deals with conditional entropy

$$H(X_n|X_1^{n-1}) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_n = x_n|X_1^{n-1} = x_1^{n-1}). \tag{9}$$

Hilberg replotted Shannon's (1951) estimates of conditional entropy for English in the double logarithmic scale and observed an approximate power-law relationship

$$H(X_n|X_1^{n-1}) \propto n^{-1+\beta}, \tag{10}$$

where $\beta \approx 0.5$ and $n \leq 100$. When extrapolated to arbitrary n, this relationship implies

$$H(X_1^n) = \sum_{m=1}^{n} H(X_m|X_1^{m-1}) \propto \int_0^n m^{-1+\beta} dm \propto n^{\beta}. \tag{11}$$

Hence we obtain a power law for the entropy rate

$$\frac{H(X_1^n)}{n} \propto n^{-1+\beta}. \tag{12}$$

Relationship (12) is the original Hilberg conjecture.

The original Hilberg conjecture is a bit far-fetched. Having derived (12), Hilberg conjectured that the entropy rate (8) of natural language is zero. This proposition seems unrealistic since it implies asymptotic determinism of human utterances. Thus it may be better to assume

$$\frac{H(X_1^n)}{n} \approx An^{-1+\beta} + h, \tag{13}$$

where constant h can be positive.

Striving for even more realism, we notice that there is no good probability distribution for texts in natural language. Hence, it seems more correct to speak of Kolmogorov complexity $K(x_1^n)$ of an individual text x_1^n rather than the entropy $H(X_1^n)$ of a random text X_1^n. Thus another plausible modification of Hilberg's conjecture reads

$$\frac{K(x_1^n)}{n} \approx An^{-1+\beta} + h. \tag{14}$$

This proposition may be called a relaxed Hilberg conjecture for individual texts. In the following, we will try to check whether (14) applies to texts in natural language. Prior to this, we will however discuss some bounds for mutual information that arise for universal codes.

4 Bounds for mutual information

It is insightful to rephrase Hilberg's conjecture using mutual information. There are three kinds of mutual information that are important for our considerations. First, the Shannon mutual information between random blocks is defined as

$$I_H(X_1^n; X_{n+1}^{2n}) = H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}). \tag{15}$$

Second, the algorithmic mutual information between individual texts is defined as

$$I_K(x_1^n; x_{n+1}^{2n}) = K(x_1^n) + K(x_{n+1}^{2n}) - K(x_1^{2n}). \tag{16}$$

Third, the mutual information based on a universal code C is

$$I_C(x_1^n; x_{n+1}^{2n}) = |C(x_1^n)| + \left|C(x_{n+1}^{2n})\right| - \left|C(x_1^{2n})\right|. \tag{17}$$

The nice feature of mutual information is that when we rephrase the modified Hilberg conjecture using this concept then the linear terms will cancel. Assuming that $H(X_1^n) \approx H(X_{n+1}^{2n})$, for the conjecture (13) we have

$$I_H(X_1^n; X_{n+1}^{2n}) \approx 2An^\beta + 2hn - A(2n)^\beta - 2hn \propto n^\beta. \tag{18}$$

Similarly, supposing that $K(x_1^n) \approx K(x_{n+1}^{2n})$, for the conjecture (14) we obtain

$$I_K(x_1^n; x_{n+1}^{2n}) \approx 2An^\beta + 2hn - A(2n)^\beta - 2hn \propto n^\beta. \tag{19}$$

In our application, we are going to estimate the algorithmic mutual information $I_K(x_1^n; x_{n+1}^{2n})$ using the code-based mutual information $I_C(x_1^n; x_{n+1}^{2n})$. It is important to know what an error is that we make by such an approximation. In determining this error the following proposition is helpful:

Lemma 1 (4) *Let a function G satisfy $\lim_{k\to\infty} G(k)/k = 0$ and $G(n) \geq 0$ for all n. Then $2G(n) - G(2n) \geq 0$ for infinitely many n.*

A nice feature of universal codes, which follows from the above lemma, is that they yield an upper bound for the Shannon mutual information. Consider a stationary process $(X_1, X_2, ...)$. By Lemma 1, from (2) and (7), we obtain

$$\sum_{x_1^n} P(X_1^{2n} = x_1^{2n}) I_C(x_1^n; x_{n+1}^{2n}) \geq I_H(X_1^n; X_{n+1}^{2n}) \tag{20}$$

for infinitely many n. For the algorithmic mutual information, we can obtain a similar statement. Consider an infinite individual text $(x_1, x_2, ...)$. Suppose plausibly that $|C(x_1^n)| \approx \left|C(x_{n+1}^{2n})\right|$, $K(x_1^n) \approx K(x_{n+1}^{2n})$, and

$$\lim_{n\to\infty} \frac{1}{n} |C(x_1^n)| = \lim_{n\to\infty} \frac{1}{n} K(x_1^n). \tag{21}$$

Then by Lemma 1, from (4) we obtain

$$I_C(x_1^n; x_1^{2n}) + c \geq I_K(x_1^n; x_1^{2n}) \tag{22}$$

for infinitely many n. Hence when $I_C(x_1^n; x_1^{2n})$ obeys a power law with a given exponent then $I_K(x_1^n; x_1^{2n})$ may only obey a power law with a smaller exponent.

The bound given in (22) is the tighter, the better the code compresses the data. Suppose that we have a code D that satisfies $|D(x_1^n)| \leq |C(x_1^n)|$ and the analogue of (21). Then by Lemma 1, we have

$$I_D(x_1^n; x_1^{2n}) \leq I_C(x_1^n; x_1^{2n}) \tag{23}$$

for infinitely many n. Hence if we look for a good estimate of algorithmic mutual information, we should use the shortest code available.

5 Empirical findings

For the sake of testing Hilberg's conjecture, we have compressed 10 texts written in English by single authors. The texts were downloaded from the Project Gutenberg[1] and are listed in Table 1. We have deleted the preambles of the text files and reduced the alphabet to 27 symbols (26 capital letters and a space), as it has been usually done in previous publication concerning the entropy of English. Subsequently, we have measured the length of the Lempel-Ziv code for exponentially growing initial text blocks.

The dependence of the compression rate on the block length is given in Figure 1, whereas the dependence of the mutual information on the double block length is given in Figure 2. Using the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm, we have fitted the following simple model for the compression rate:

$$\frac{|C(x_1^n)|}{n} \approx 6.22 n^{-1+0.949} \text{ [bpc]}. \tag{24}$$

[1] `http://www.gutenberg.org/`

Title	Author
First Folio/35 Plays	W. Shakespeare
Critical & Historical Essays	T. B. Macaulay
The Complete Memoirs	J. Casanova
Memoirs of Comtesse du Barry	E. Lamothe-Langon
The Descent of Man	C. Darwin
Gulliver's Travels	J. Swift
The Mysterious Island	J. Verne
Mark Twain, a Biography	A. B. Paine
The Journal to Stella	J. Swift
Life of William Carey	G. Smith

Table 1: The selection of compressed texts.

From formula (24), we derive mutual information

$$I_C(x_1^n; x_{n+1}^{2n}) \approx 0.432n^{0.949} \text{ [bits]}. \tag{25}$$

In Figures 1 and 2, we can observe that both models fit the data very well.

It may be somewhat surprising that model (24) fits so well although it contains no constant term $h > 0$ supposed in conjecture (14). We know, however, from independent studies that the asymptotic entropy rate h for English is less than 1.25 bpc [2]. In contrast, the lowest compression rate that we observe in Figures 1 is about 3.0 bpc. Thus a constant term of the order of 1.25 bpc cannot be reliably identified in the considered data.

Relationships (22) and (25) suggest that this bound holds for the algorithmic mutual information of texts in English:

$$I_K(x_1^n; x_{n+1}^{2n}) \leq 0.432n^{0.949} + c \text{ [bits]}. \tag{26}$$

The above relationship does not exclude Hilberg's conjecture with a very high exponent β.

6 Conclusion

In this paper, we have first presented an approach how to understand Hilberg's conjecture using Kolmogorov complexity and algorithmic mutual information. Putting Hilberg's conjecture in this setting escapes the problem of deciding what is an appropriate probability distribution for human language production. In the second turn, we have tried to verify Hilberg's conjecture using the Lempel-Ziv code. Our findings do not exclude Hilberg's conjecture with an exponent β close to 1. However, if we used a better universal code than the Lempel-Ziv code, such as the PPM code, then we might obtain a tighter bound for the exponent. We leave this problem for the future research.

References

[1] Behr, F., Fossum, V., Mitzenmacher, M., Xiao, D.: Estimating and comparing entropy across written natural languages using PPM compression. Tech. Rep. TR 12-02, Harvard University (2002)

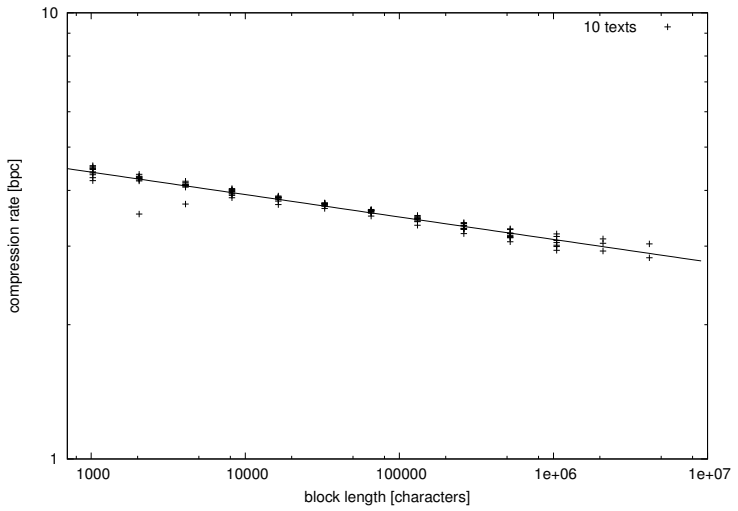

Figure 1: Compression rate vs. block length. The solid line corresponds to the curve (24).

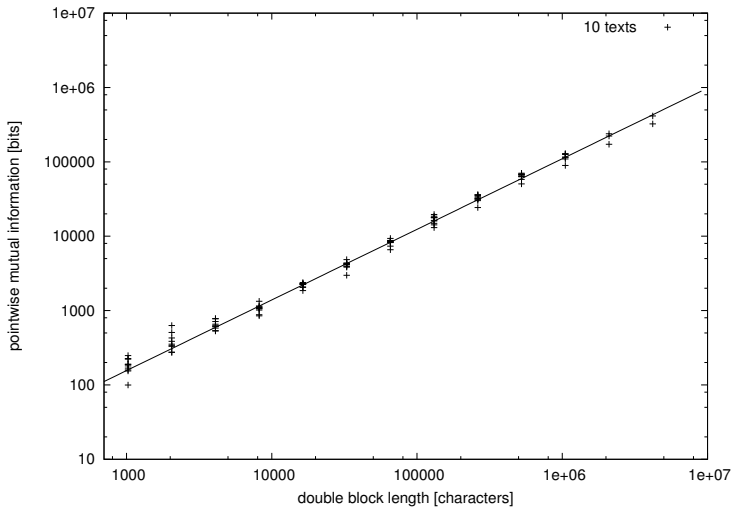

Figure 2: Mutual information vs. double block length. The solid line corresponds to the curve (25).

[2] Cover, T.M., King, R.C.: A convergent gambling estimate of the entropy of English. IEEE Transactions on Information Theory 24, 413–421 (1978)

[3] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd ed. New York: John Wiley (2006)

[4] Dębowski, Ł.: On the vocabulary of grammar-based codes and the logical consistency of texts. IEEE Transactions on Information Theory 57, 4589–4599 (2011)

[5] Dębowski, Ł.: Maximal lengths of repeat in English prose. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.) Synergetic Linguistics. Text and Language as Dynamic System, pp. 23–30. Wien: Praesens Verlag (2012)

[6] Herdan, G.: Quantitative Linguistics. London: Butterworths (1964)

[7] Hilberg, W.: Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? Frequenz 44, 243–248 (1990)

[8] Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission 1(1), 1–7 (1965)

[9] Mahoney, M.V.: Text compression as a test for artificial intelligence. In: Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference, AAAI'99/IAAI'99, p. 970 (1999)

[10] Shannon, C.: Prediction and entropy of printed English. Bell System Technical Journal 30, 50–64 (1951)

[11] Teahan, W.J., Cleary, J.G.: The entropy of English using PPM-based models. In: Proceedings of the Conference on Data Compression, DCC'96. pp. 53–62. Washington, DC: IEEE Computer Society (1996)

[12] Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Transactions on Information Theory 23, 337–343 (1977)