A New Universal Code Helps to Distinguish Natural Language from Random Texts

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

Abstract
Using a new universal distribution called switch distribution, we reveal a prominent statistical difference between a text in natural language and its unigram version. For the text in natural language, the cross mutual information grows as a power law, whereas for the unigram text, it grows logarithmically. In this way, we corroborate Hilberg’s conjecture and disprove an alternative hypothesis that texts in natural language are generated by the unigram model.

Keywords: universal coding, switch distribution, Hilberg’s conjecture, unigram texts

1 Introduction
G. K. Zipf supposed that texts in natural language differ statistically from the output of producing characters at random (Zipf, 1965, p. 187), or as we would say in modern terms, from unigram texts. With the advent of computational tools, this difference can be revealed using various experimental methods. For example, although both texts in natural language and unigram texts obey some versions of Zipf’s law (Mandelbrot, 1954; Miller, 1957), some difference between texts in natural language and unigram texts can be discerned by investigating merely the rank-frequency distribution of words (Ferrer-i-Cancho and Elvevåg, 2010).

In this paper we wish to demonstrate a more prominent difference between random texts and texts in natural language which can be detected by means of universal coding. This experimental setup is closely related to Hilberg’s conjecture (Hilberg, 1990), an important hypothesis concerning the entropy of natural language. According to this hypothesis, natural language production forms a stationary stochastic process \((X_i)_{i=-\infty}^{\infty}\) on some probability space \((\Omega, \mathcal{F}, P)\) with blocks of consecutive symbols being denoted as \(X^n = (X_1, X_2, \ldots, X_n)\), whereas the pointwise entropy of text blocks of length \(n\), \(n\) denoted \(H_P(n) = -\log P(X^n)\), satisfies

\[
H_P(n) \approx Bn^\beta + hn,
\]

(1)
where $\beta \approx 0.5$ and $h \approx 0$ (Hilberg, 1990). This property may distinguish texts in natural language from unigram texts. Assuming that the distribution of characters is a vector of unknown random parameters of a unigram text, we obtain for this text that

$$H^P(n) \approx B \log n + hn,$$

where $h > 0$ and $B$ is proportional to the number of parameters, cf. Grünwald (2007).

In both expressions (1) and (2) the value of entropy is asymptotically dominated by the linear term $hn$. To make the difference between (1) and (2) more prominent, we may consider mutual information between adjacent blocks, $I^P(n) = 2H^P(n) - H^P(2n)$. In this way we obtain

$$I^P(n) \propto n^\beta$$

if Hilberg’s conjecture is satisfied, whereas

$$I^P(n) \propto \log n$$

for unigram texts. Relationship (3), which we call the relaxed Hilberg conjecture, was initially investigated by physicists interested in complex systems (Ebeling and Nicolis, 1991, 1992; Ebeling and Pöschel, 1994; Bialek et al., 2001b,a; Crutchfield and Feldman, 2003) but later interesting linguistic interpretations were provided by Debowksi (2006, 2011). There are in particular mathematical connections between the relaxed Hilberg conjecture (3) and various forms of Zipf’s and Herdan’s laws (Herdan, 1964).

Testing Hilberg’s conjecture is difficult. What we need are estimates of entropy $H^P(n)$ or mutual information $I^P(n)$ for block lengths $n$ varying over a large subset. It is quite costly or impossible to obtain such estimates using the guessing method (Shannon, 1951) or the gambling method (Cover and King, 1978). As an alternative, we may consider universal coding or universal distributions. For another probability distribution $Q$, let us denote the pointwise cross entropy $H^Q(n) = -\log Q(X^n_1)$ and the pointwise cross mutual information $I^Q(n) = 2H^Q(n) - H^Q(2n)$. By Barron’s “no hypercompression” inequality (Barron, 1985, Theorem 3.1, Grünwald, 2007, p. 103), the cross entropy is greater than the entropy plus a logarithmic term,

$$H^Q(n) \geq H^P(n) - 2 \log n$$

for almost all $n$ almost surely. Moreover, the defining property of a universal distribution $Q$ is that the compression rate $H^Q(n)/n$ tends to the entropy rate for the text length tending to infinity,

$$\lim_{n \to \infty} \frac{H^Q(n)}{n} = \lim_{n \to \infty} \frac{H^P(n)}{n}$$

(Cover and Thomas, 1991). Combining (5) and (6), we obtain that the cross mutual information is greater than the true mutual information,
\[ I^Q(n) \geq I^P(n) - 2 \log n \]  

(7)

for infinitely many \( n \) (Dębowski, 2011, Lemma 1). In this way, the difference of lengths of a universal code \( I^Q(n) \) is some estimate of mutual information \( I^P(n) \). In principle this estimate might be used for testing Hilberg’s conjecture but the problem is that we ignore how large the difference \( I^Q(n) - I^P(n) + 2 \log n \) is.

Dębowski (2013a) tried to test Hilberg’s conjecture using the Lempel-Ziv code (Ziv and Lempel, 1977). This is the oldest known example of a universal code. In fact, the length of the Lempel-Ziv code for texts in natural language satisfies

\[ I^Q(n) \propto n^\beta \]  

(8)

with \( \beta \approx 0.9 \) for text lengths in the range of \( n \in (10^3, 10^7) \) characters. The problem is, however, that the difference of lengths of the Lempel-Ziv code \( I^Q(n) \) for a unigram text is very similar, as we will show in the experimental part of this paper, cf. a theoretical result by Louchard and Szpankowski (1997). Hence the Lempel-Ziv code cannot be used for discriminating between texts in natural language and unigram texts. This does not mean that we cannot discriminate between natural language and unigram texts by universal coding in principle, but rather another universal code should be used for that purpose.

Actually, we have introduced recently some prospective universal code, called switch distribution (Dębowski, 2013b). The switch distribution is a development of an idea by van Erven et al. (2007). It is a generic probability distribution for data prediction. Formally, it is a mixture of infinitely many Markov chains of all orders but it is effectively computable and proved to be a universal code. For the exact formula, which is a bit complicated, we refer to Dębowski (2013b). As we will show in the experimental part of this paper, for unigram texts the length of the switch distribution satisfies

\[ I^Q(n) \propto \log n \]  

(9)

unlike the Lempel-Ziv code, whereas the length of the switch distribution satisfies (8) for texts in natural language, like for the Lempel-Ziv code. In this way, we can discriminate between natural language and unigram texts using the switch distribution. Our observation also makes Hilberg’s conjecture more likely.

The remaining part of the paper consists of two sections. In Section 2, we present the experimental data, whereas in Section 3 we offer concluding remarks.

2 Experimental data
In this section we experimentally investigate the length of the Lempel-Ziv code and the switch distribution for a text in natural language and a unigram text. The considered text in natural language is “20,000 Leagues under the Sea” by Jules Verne, whereas the other text is the unigram model of this novel. The experimental data with regression functions are given in Figures 1, 2, 3, and 4.

Visually, in case of the Lempel-Ziv code, we observe no substantial difference between the two considered texts. For both texts, the compression rate $H^Q(n)/n$ decreases hyperbolically, viz. Figure 1, whereas the difference of code lengths $I^Q(n)$ grows like a power law, viz. Figure 2. If we apply the plain switch distribution, however, there arises a huge difference. In Figure 3, the compression rate for the text in natural language decreases hyperbolically, whereas the compression rate for the unigram text stabilizes. Moreover, in Figure 4, the difference of code lengths for the text in natural language grows like a power law, whereas for the unigram text it grows logarithmically.

The data points can be approximated by the following least squares regression functions, where the values after the ± sign are standard errors. For the Lempel-Ziv code and the text in natural language, we have

$$ I^Q(n) = (0.64 \pm 0.17)n^{(0.936\pm0.002)}, \quad (10) $$

whereas for the Lempel-Ziv code and the unigram text, we obtain

$$ I^Q(n) = (1.31 \pm 0.05)n^{(0.832\pm0.003)}, \quad (11) $$

viz. Figure 2. In contrast, for the switch distribution and the text in natural language, we obtain

$$ I^Q(n) = (0.67 \pm 0.10)n^{(0.898\pm0.010)}, \quad (12) $$

whereas for the switch distribution and the unigram text, we have

$$ I^Q(n) = (74 \pm 4)\log(0.027 \pm 0.006)n, \quad (13) $$

except for the three last data points, viz. Figure 4.

The three last data points for the switch distribution and the unigram text are probably outliers, caused by some unknown numerical errors, since mathematical theory predicts that the cross mutual information for the switch distribution on a unigram text follows relationship (9) asymptotically. This theoretical result stems from the fact that the switch distribution is a mixture of Markov chains of all orders, whereas for the mixture of Markov chains of zeroth order we obtain the scaling (9) (Grünwald, 2007). In spite of the mentioned numerical error, using the switch distribution, we can still distinguish the text in natural language from the unigram text. As we have stated in the introduction, this possibility supports Hilberg’s conjecture.

Another interesting result is that we can reject with an extremely low p-value the
hypothesis that the text in natural language was generated by the unigram model. For that sake we use a stronger form of Barron’s “no hypercompression” inequality. Suppose we have two probability distributions $P$ and $Q$. The stronger form of the Barron inequality states that the probability of sufficiently long random data $X^n_t$ according to distribution $P$ is very low if only we can compress it better using an alternative distribution $Q$. Namely, we have

$$P(H^Q(n) \leq H^P(n) - m) \leq 2^{-m}$$

(14)

(Barron, 1985, Theorem 3.1, Grünwald, 2007, p. 103). In our case $P$ is the unigram model, $Q$ is the switch distribution, whereas $X^n_t$ is the text of “20,000 Leagues under the Sea” with length $n = 524,288$ characters. For the unigram model we obtain compression rate $H^P(n)/n = 4.4481$ bpc (bits per character), whereas for the switch distribution we have $H^Q(n)/n = 2.3018$ bpc. Hence the probability that the text can be so compressed by the switch distribution if it were generated by the unigram model is less than

$$2^{-524,288(4.4481-2.3018)} \leq 2^{-1,000,000} .$$

(15)

Thus the unigram model should be rejected. We think that Barron’s inequality can be used in a similar fashion for disproving other simple probabilistic hypotheses about natural language.

3 Conclusion

In this paper, using a new universal distribution called switch distribution, we have shown a prominent statistical difference between a text in natural language and its unigram version. The difference consists in a different growth rate of cross mutual information. Namely, for the text in natural language, the cross mutual information grows as a power law, whereas for the unigram text, it grows logarithmically. This observation corroborates Hilberg’s conjecture, an important hypothesis concerning natural language, and disproves the alternative hypothesis that texts in natural language can be generated by the unigram model. Further investigation is needed to illuminate why for the text in natural language we observe the power law exponent $\beta \approx 0.9$ whereas the data analyzed by Hilberg (1990) suggest that $\beta \approx 0.5$.

References


Louchard, G., Szpankowski, W., 1997. On the average redundancy rate of the Lempel-


Figure 1: Compression rate $H^Q(n)/n$ vs. block length $n$ for the Lempel-Ziv code.

Figure 2: Pointwise cross mutual information $I^Q(n)$ vs. block length $n$ for the Lempel-Ziv code. The lines are the regression functions.
Figure 3: Compression rate $H^Q(n)/n$ vs. block length $n$ for the switch distribution.

Figure 4: Pointwise cross mutual information $I^Q(n)$ vs. block length $n$ for the switch distribution. The lines are the regression functions.