Manuscript

ENTROPY ANALYSIS OF N-GRAMS AND ESTIMATION OF THE NUMBER OF MEANINGFUL LANGUAGE TEXTS. CYBER SECURITY APPLICATIONS

February 11, 2021

Anastasia Malashina
Computer Security Department
HSE University, Moscow Russia
amalashina@hse.ru

Abstract We estimate the n-gram entropies of English-language texts, using dictionaries and taking into account punctuation, and find a heuristic method for estimating the marginal entropy. We propose a method for evaluating the coverage of empirically generated dictionaries and an approach to address the disadvantage of low coverage. In addition, we compare the probability of obtaining a meaningful text by directly iterating through all possible n-grams of the alphabet and conclude that this is only possible for very short text segments.

Keywords: n-gram entropy, n-gram dictionaries, coverage, meaningful texts, information theory.

1 Introduction

Entropy is the basis of the information-theoretic approach to information security. It is a degree of uncertainty. The data with maximum entropy is completely random, and no patterns can be established. For low-entropy data, we are able to predict the following generated values. The level of chaos in the data can be calculated using the entropy values of the system. The higher the entropy, the greater the uncertainty and unpredictability, the more chaotic the system.

Text is also a system that has entropy. Moreover, natural language texts have entropy significantly lower than the maximum entropy of the alphabet. In turn, a random non-meaningful set of characters has the maximum possible entropy in a given alphabet. The entropy index can be used for automatic recognition of the text meaningfulness when searching through various decryption options or when dictionary attack [1].

In addition, entropy can be used in the keyless recovery method of encrypted information. If we divide an encrypted message into discrete segments of a fixed length, the entropy value determines how many meaningful text recovery options there are for each such segment of the message. Since the number of meaningful texts in the language is significantly less than arbitrary (random) ones, this approach critically reduces the complexity of decryption compared to a brute force attack [2].

There are various methods for determining the entropy of a text or its individual segments, called n-grams. The most popular of them is the Shannon method, based on the representation of the text by a Markov chain of depth n [3]. In this paper, we propose to use a dictionary-based method for determining the entropy of n-grams, combining the Kolmogorov's combinatorial approach and a corollary of the Shannon's second theorem, which gives an asymptotic estimate of the number of meaningful texts. Moreover, we propose a theoretical approach to estimating the coverage of the created n-gram dictionaries and a method for correcting the accuracy of its volume.

We explore texts in English collected from various web pages based on corpus [4], considering an extended alphabet that includes the simplest punctuation marks. The aim of the study is to evaluate the entropy of short-length texts and to extend the results obtained to long texts. Based on the entropy data, we theoretically estimate the number of long meaningful texts in a language for which an empirical estimate is impossible.

The structure of the paper is as follows: in Section 2, we describe the corpus of analyzed texts and the preprocessing of the corpus, and in Section 3, the methodology used for n-gram dictionaries and coverage, as well as the estimation of n-gram entropies. Section 4 presents and discusses the results of our analysis. Section 5 summarizes the main conclusions of this article.

2 Corpus Description

The corpus we analyze is based on text samples from the iWeb corpus of English language [4] and contains about 100 million characters collected from web pages.

The corpus we created must meet two main criteria: completeness and representativeness. The completeness is determined by the coverage of this corpus. Optimization of the coverage depends on the problems considered [5]. First, the coverage depends on the text corpus volume that is used for compiling dictionaries. But as the corpus volume increases, this dependence becomes less pronounced, so data can be extrapolated for further volumes of dictionaries [6]. For example, for English, the growth of the dictionary volume slows down significantly when the corpus size is from 30 to 50 million words. Second, the optimal size of the corpus depends on the sources and novelty of the data [7]. In General, the corpus is considered saturated when the sharp growth of new words stops with an increase in the corpus volume [5]. There is no metadata in the corpus, since it is not essential for further use of this corpus.

Based on the text corpus collected and in accordance with the n-gram language model, we create dictionaries for further research.

2.1 Data preprocessing

To increase the coverage and relevance of n-gram dictionaries, we restrict the size of the alphabet. Therefore, the corpus created goes through a filtering process. In addition, we delete errors and typos from the text to minimize the probability that type II errors will appear: we assume that only meaningful texts are represented in the n-gram vocabularies.

In general, the normalization process consists of the following steps:
1) deleting HTML tags,

2) recoding,
3) filtering the text (deleting all characters except “a-z”, “.”, “,”,” “, cast to lowercase),
4) removing double spaces, repeating dots and commas, and spaces before dots and commas,
5) deleting errors and typos.

Thus, the alphabet power of our corpus is 29 characters. We consider it as a simple extension of the Latin alphabet including punctuation.

3 Methodology

3.1 N-gram vocabularies

Within the n-gram model of the language, we generate the dictionaries. The dictionary units are n-grams. We consider an n-gram as a sequence of n characters. N-grams are selected from the text with chaining: for the next n-gram, we shift to the right by one character. An example of this process is shown in Figure 1.

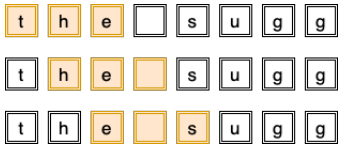

Fig. 1 The process of trigram selection with chaining.

The process of a dictionary creation consists of extracting all n-grams from the corpus in the list, deleting duplicate n-grams and sorting.

We fix the total number of n-grams in the corpus, and the vocabulary volume, that is, the number of non-repeating n-grams in the corpus. Besides, we find the number of unique n-grams in the dictionary, i.e. units of vocabulary, occurring only once.

We assume that the optimal length of a text segment to examine is between 10 and 25 characters. Therefore, we generate the n-gram dictionaries of 10, 15, 20, and 25 characters. We process about 100 million n-grams. Different lengths of n-grams help us study how the characteristics of dictionaries change with increasing length of the text segment.

The compiled n-gram dictionaries form the basis of our methodology for calculating the entropy values. We assume that the created dictionaries are a tool for automatically distinguishing meaningful and random n-grams of a language.

3.2 Coverage and vocabulary resizing

Since dictionaries are compiled on the corpus of limited length, their coverage is incomplete. This means that not all meaningful n-grams of the language are included in this dictionary. That is, type I errors are possible, when a meaningful n-gram that is not present in the dictionary is discarded as random. This situation, for example, is possible when organizing a dictionary attack. Therefore, it is necessary to evaluate the coverage of the dictionaries created, that is, to estimate what proportion of meaningful n-grams of the language our dictionaries cover.

In this study, we propose a theoretical estimate of the coverage that is independent on empirical tests. The theoretical assessment is based on the number of unique n-grams in the dictionary, i.e. units of vocabulary, occurring only once:

$$coverage = (1 - \frac{k}{K}) \cdot 100\%, \tag{1}$$

where K is the initial dictionary volume, and k is the number of n-grams that occur once.

If the coverage of the initial dictionary is low, then the empirical estimates derived from it may not be accurate enough. Therefore, it is necessary to recalculate the volume of the empirical dictionary to bring it closer to the real one. We propose the approach to resizing of dictionaries.

Let us obtain a dictionary of K units with k elements occurring once. Then $1 - \frac{k}{K}$ is a fraction of repetitive elements in the dictionary, and $(1 - \frac{k}{K}) \cdot K$ is the number of duplicate elements in the initial dictionary. Since it is compiled on the corpus of limited volume, this dictionary does not have full coverage. Obviously, $(1 - \frac{k}{K}) \cdot K < K$. It is known that up to a certain point, the number of new n-grams in the dictionary grows at a linear rate [7]. To get a dictionary consisting of K repeated elements, we need to increase its volume by $\frac{1}{1-\frac{k}{K}}$ times.

Thus, a new dictionary with volume

$$\tilde{K} = \frac{K}{1 - \frac{k}{K}} \tag{2}$$

contains about K repeated elements. The new volume of the dictionary compensates for the lack of coverage of the original one.

Thus, $\tilde{K}$ is a theoretical estimate of the n-gram dictionary volume, obtained on the basis of an empirical dictionary.

The theoretical n-gram vocabulary function is

$$\tilde{K}(n) = 2^{H \cdot n}. \tag{3}$$

This estimation [8] is based on the language entropy H.

3.3 N-gram entropy

We assume that the volume of the dictionary obtained is the number of meaningful n-grams in the language. This means that we consider all out of vocabulary n-grams as random texts. Based on this assumption, we can estimate the entropy of n-grams.

Within the n-gram language model, the text is the implementation of independent tests, the results of which are the n-grams of the corresponding natural language. Then the entropy per sign of the text is estimated as $\frac{\hat{H}_n}{n}$, where $\hat{H}_n$ is the entropy of a random source, where the outcomes are n-grams. The limit of this relation is the language entropy [3].

The entropy of a random source may be calculated in a classical way [3]

$$\hat{H}_n = -\sum_{(a_{i_1},\ldots,a_{i_n})} p(a_{i_1},\ldots,a_{i_n}) \cdot \log_2 p(a_{i_1},\ldots,a_{i_n}) \qquad (4)$$

where $p(a_{i_1},a_{i_2},\ldots,a_{i_n})$ is the probability of the i-th n-gram.

To avoid calculating n-gram probabilities, we propose a simple empirical way for calculating entropy based on the dictionary volume. This idea is based on the Kolmogorov's combinatorial method and the Shannon's second theorem [9].

Since the number of all distinct n-grams in the alphabet with power A is estimated as $A^n = 2^{n\cdot\log_2 n} > \tilde{K}$ is greater than the number of meaningful n-grams in the same language, then there is a value $\hat{H}_n$, such that $\tilde{K} = 2^{n\cdot\hat{H}_n}$, where $\hat{H}_n > \log_2 A$ [3].

Let the dictionary volume $\tilde{K}$ be the number of meaningful n-grams in the language. Then $\hat{H}_n = \log_2 \tilde{K}$. Value $\hat{H}_n$ is the entropy of n-grams (bits). Therefore, the larger the dictionary coverage, the more accurate the entropy estimate.

The entropy of n-grams per character (bits/symbol):

$$H_n = \frac{\log_2 \tilde{K}}{n} \qquad (5)$$

4 Results and Discussion

In Figure 2, we present the results of the estimation of the n-gram vocabularies for short length texts. As said in Section 3.1, the shown values are the empirically obtained number of total n-grams and n-grams occurred only once.

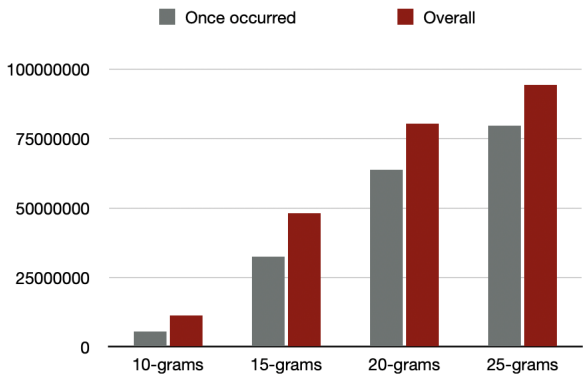

Fig. 2 N-gram vocabularies for shortlength texts.

Table 1 Initial coverage and new size of vocabulary

Text length	Initial coverage, %	Theoretical vocabulary
10	51.35	21912941
15	32.33	148511061
20	20.84	385966128
25	15.59	605647799

One can notice that the number of n-grams that occur only once increases significantly with the length of the text segment. Regarding the total number of n-grams, we have found that the growth rate of the dictionary volume is less than linear one.

Based on the number of n-grams occurred only once, we can estimate the coverage of empirically obtained dictionaries. Since the coverage of the source dictionaries is insufficient, it is necessary to recalculate the volumes of the n-gram dictionaries using the methodology proposed in Section 3.1. The coverage values and volumes of new dictionaries are shown in Table 1.

As expected, the new dictionaries correspond to a more complete coverage and are used in subsequent stages of the study.

To investigate how the volume of dictionaries changes depending on the corpus size, we have found an interpolation function. In Figure 3, we show this interpolation model for 10-grams. We can see that the growth rate of dictionaries is below the linear dependence is the closest to the square root function.

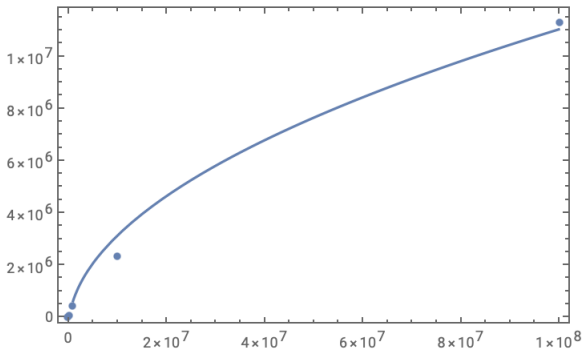

Fig. 3 10-gram dictionary interpolation function.

Equation 6 describes the interpolation function for 10-grams.

$$-1.14697 \cdot 10^6 + 1230.21\sqrt{x} \qquad (6)$$

For other values of n, the interpolation function remains the same with a slight change in the coefficients. Using this model, we can predict some subsequent values. For exam-

ple, for a corpus of 300 million characters, we expect a dictionary of 20 million 10-grams, and for a corpus of 700 million characters, we could expect a dictionary of 30 million 10-grams.

4.1 Distribution of n-gram entropy

It is well-known the amount of information transmitted by a single n-gram increases with the length of the segment. To determine the average amount of information per character, i.e. the specific information content of the source, we need to divide this number by n. With unlimited growth, the approximate equality will turn into the exact one. The result is an asymptotic relation.

Using the approach presented in Section 3.2, we have determined the entropy of short-length texts based on the volume of the original empirical dictionary K and the theoretical one $\tilde{K}$. In Figure 4, we can see that the specific entropy of the source (text) decreases with increasing length of the n-gram.

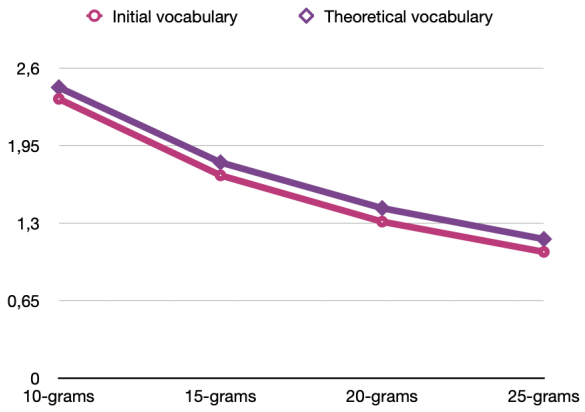

Fig. 4 Entropy per character for short length texts.

The entropy per symbol means that it takes H_n bits of information to determine the $n+1$ character of the text. The more information we know, the less uncertainty there is about the next character in the text. This fact explains the decreasing nature of the specific entropy function.

However, as the text length increases, the rate of entropy decrease slows down. For example, the difference between H_{10} and H_{15} is only about 0.63 bits. It means that if we know the first 10 characters of a substring of length 15, there is little uncertainty about the remaining 5 characters.

It is important to note that we have considered the alphabet extended, so the resulting n-gram entropy values differ from the known values for English.

With the growth of n, the value of H_n decreases to some n and with further growth almost does not change, that is, it reaches a certain limit, called the entropy of the language. However, our n-gram model is based on a finite corpus of text samples, so estimating the entropy rate for large values of n gives implausibly low information rates. In other words, as the value of n increases, the experimental estimates of entropy per symbol tend to 0. Indeed, as the model order increases, the number of n-grams samples decreases, so that for very large values of n, knowledge of the first $n-1$ letters of the text uniquely identifies the text in question, that is, the n-letter is pre-determined.

Extrapolating these results to large values of n is difficult, because the shape of this sequence of values is generally unknown, except that it is positively decreasing. To obtain the ultimate entropy from this set of measurements, we construct a model of sequential estimates.

We have assumed that the sequence of entropy values obeys a linear recurrence relation:

$$F_n - F_{n+1} = k \cdot (F_{n+1} - F_{n+2}) \tag{7}$$

with initial conditions $F_0 = H_{10}$ and $F_1 = H_{15}$, where $F_n = H_{5s}$.

Then the k that yields the best approximation is $k = 0,62$.

In Figure 5, we present the extrapolation results of entropy per character for the initial and theoretical vocabularies $\tilde{K}$.

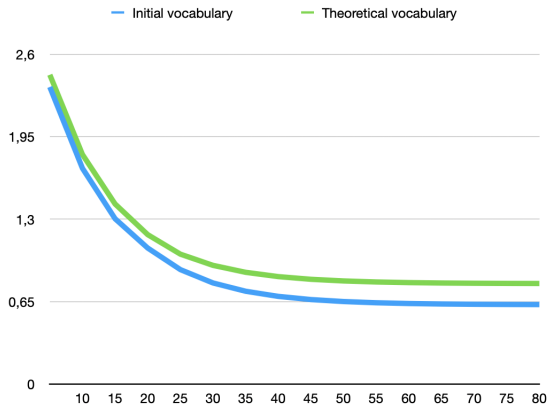

Fig. 5 Extrapolated entropy rate values.

This extrapolation model is heuristic, but it is sufficient to solve the problems set in the paper. As we can see on the graph, the limit entropy rate is 0.8 bits per character for the theoretical vocabulary. Therefore, we can estimate the entropy H_n as the limit entropy for a long text.

4.2 Number of meaningful texts

Using the entropy values obtained and the extrapolation model constructed, we have estimated the number of meaningful

Table 2 Number of meaningful texts

Length of text segment	Number of meaningful texts	Relative part
30	$1.8 \cdot 10^{9}$	$2.4 \cdot 10^{-35}$
50	$3 \cdot 10^{12}$	$2.3 \cdot 10^{-61}$
100	$1.2 \cdot 10^{24}$	$0.7 \cdot 10^{-122}$
300	$1.8 \cdot 10^{72}$	$0.3 \cdot 10^{-366}$
500	$2.6 \cdot 10^{120}$	$1.7 \cdot 10^{-611}$
1000	$6.7 \cdot 10^{240}$	$2.7 \cdot 10^{-1222}$

texts among all texts of fixed length. The results are shown in Table 2.

Meaningful texts are n-grams that exist in the language considered. N-grams that are not meaningful are a random set of n characters.

The last finding is the proportion of meaningful n-grams among the total number of n-grams. In other words, it is the probability of finding a meaningful text with a random sample from among all possible n-grams. As expected, this probability decreases with increasing text length. As the text length increases, the probability of finding a meaningful n-gram critically decreases. This confirms the previously stated hypothesis that to recover individual parts of the encrypted text, it is worth considering n-grams of short length.

5 Conclusion

In this paper, we have estimated the n-gram entropies of natural language texts and examined the number of meaningful texts in English. Most of the previous studies on the n-grams entropy did not take into account punctuation, so the values obtained in this paper have eliminated this gap. We have found that the empirical method of generating dictionaries can lead to significant type I errors in estimating the number of meaningful n-grams due to low coverage. We have eliminated this drawback by offering a method for refining the theoretical volume.

The entropy of the text per character decreases positively with the growth of the n-gram length. This can be explained by the fact that as the length of the known text increases, the uncertainty of the next character decreases. However, starting from a certain n, the entropy values almost do not change, reaching a certain plateau, called the entropy of the language. By extrapolating the data with a linear recurrent sequence, we have heuristically determined the limiting entropy of our corpus, which is 0.8 bits per character.

The found limit value of entropy allowed us to estimate the number of meaningful long-length n-grams, so it is almost impossible to do empirically. The probability of finding a meaningful text among all possible sets of n-grams for large n is catastrophically small. This result confirmed our assumption that it is advisable to use short n-grams to recover information using the information-theoretic approach.

References

1. A. Jaglom, I. Jaglom, Moscow: Science (in Russ.) (1973)
2. D. Florencio, C. Herley, in *Proceedings of the 16th international conference on World Wide Web* (2007), pp. 657–666
3. C.E. Shannon, The Bell system technical journal **27**(3), 379 (1948)
4. M. Davies, *iWeb: The 14 Billion Word Web Corpus* (Brigham Young University, 2018)
5. R. Rosenfeld, in *Fourth European Conference on Speech Communication and Technology* (1995)
6. J. Bellegarda, (2001)
7. L. Chase, R. Rosenfeld, W. Ward, in *Third International Conference on Spoken Language Processing* (1994)
8. J.L. Massey, in *Proceedings of 1994 IEEE International Symposium on Information Theory* (IEEE, 1994), p. 204
9. D. Sayre, Acta Crystallographica **5**(6), 843 (1952)