Baltic J. Modern Computing, Vol. 9 (2030), No. 3, pp. 363–376
https://doi.org/10.22364/bjmc.2021.9.3.09

The Combinatorial Analysis of n-Gram Dictionaries, Coverage and Information Entropy based on the Web Corpus of English

Anastasia MALASHINA

Department of Computer Security, HSE University, Moscow, Russia

amalashina@hse.ru

Abstract. We research n-gram dictionaries and estimate its coverage and entropy based on the web corpus of English. We consider a method for estimating the coverage of empirically generated dictionaries and an approach to address the disadvantage of low coverage. Based on the ideas of Kolmogorov's combinatorial approach, we estimate the n-gram entropy of the English language and use mathematical extrapolation to approximate the marginal entropy. In addition, we approximate the number of all possible legal n-grams in the English language for high order of n-grams.

Keywords: n-gram entropy, n-gram dictionaries, coverage

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 set of characters has the maximum possible entropy in a given alphabet. The entropy index can be used for automatic recognition if the text is legal in the language when searching through various decryption options or when dictionary attack, as described by Jaglom and Jaglom (1973).

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 possible text recovery options there are for each such segment of the message. Since the number of existing 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. A similar approach was used for passwords by Florencio and Herley (2007).

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 (1948) method. Using representation of the text by a Markov chain of depth n, it is possible to approximately estimate the probabilities of n-grams. In this paper, we propose to use a dictionary-based method for determining the entropy of n-grams, whose ideas go back to the Kolmogorov (1993) approach. Moreover, we propose a theoretical method to estimating the coverage of the created n-gram dictionaries and a approach for correcting the accuracy of its volume.

We explore texts in English collected from various web pages, measured by n-gram language model, considering an extended alphabet that includes the simplest punctuation marks. The aim of the study is to evaluate the entropy of short-length n-grams based on the corpus and to extend the results obtained to long n-grams. Using the entropy data, we theoretically estimate the approximate number of long legal n-grams in the 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 Related works

The n-gram model is one of the most widely used model for natural language modeling. These n-grams are related to Markov models that estimate next symbol from a fixed set of previous symbols. For n-grams its probabilities can be estimated by counting in a corpus and normalizing via the maximum likelihood estimate. If the numerical estimates for the n-gram model are determined based on the same corpus in which they appear, then such an estimate is considered intrinsic according to Jurafsky and Martin (2009).

There are various algorithms used to improve the accuracy of determining the n-gram probabilities and smoothing the coverage. These algorithms are based on counting lower-order n-grams by backtracking or interpolation.

The problem is that any empirical textual material is limited and *a priori* does not include all the existing n-grams of the language (such n-grams are called out-of-vocabulary). Then the coverage is associated with an estimate of the percentage of such elements, that is, the OOV rate. In problems of speech recognition and machine translation, the problem of out-of-vocabulary elements is often solved by using closed dictionaries, that is, the existence of OOV n-grams is ignored.

The problem of estimation and optimization of coverage is periodically considered in different subtasks. The problem of n-gram coverage often arises in machine learning

and machine translation tasks. Methods for increasing the coverage of n-grams based on the alignment entropy is proposed in Poncelas et al. (2017), but this approach uses a parallel pair of text corpora and is not applicable to the self-assessment of the coverage of a single corpus.

Rosenfeld (1995) ascertained that optimization of the coverage depends on the problems considered. 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 according to Bellegarda (2001). 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 according to Chase et al. (1994). In General, Rosenfeld (1995) states the corpus is considered saturated when the sharp growth of new words stops with an increase in the corpus volume. There is no metadata in the corpus, since it is not essential for further use of this corpus.

Markov models are often used as an approximate models of natural language. As described by Cover and King (1978), the Markov process is stationary, that is the probability distribution for n-grams at time t is the same as the probability distribution at time $t + 1$, but any natural language is not stationary, since the probability of upcoming n-grams can depend on events that were arbitrarily distant and time-dependent. Thus, these statistical models only give an approximation to the correct natural language distributions and entropies.

Despite the existence of other models, for example described by Chomsky (1956), many studies of natural languages, and in particular English, use the approximation of the text by the stationary Markov process. For example, in the papers of Calin (2020), Hahn and Sivley (2011), Yadav et al. (2010), Guerrero (2009) the Markov process is used to simulate a natural language text. Since the accuracy of approximating a natural language text using the Markov model decreases significantly with an increase in the order of n-gram, related studies mainly investigate the entropy of short-length texts. For instance, the research of Guerrero (2009) explores models of n-grams only to order 15. Therefore, there is some gap in research related to the high-order n-gram entropy.

Kolmogorov (1993) proposed an alternative combinatorial approach to the study of the entropy of language. Such a purely combinatorial approach evaluates the flexibility of the language, that is, it gives an estimate of the number of text continuations with a fixed dictionary and phrase construction rules. This method tends to overestimate the real values of the language entropy, since any meaningful texts in natural language are subject not only to grammatical rules, but also have some content constraints. Nevertheless, there are areas of research, such as White (1967) paper, in which a certain vocabulary is fixed initially. Then, for such closed dictionary systems, the combinatorial approach can give fairly accurate entropy estimates.

Regarding the study of the marginal entropy of the English language, at different times there are a number of studies that consider different approaches to the assessment of the entropy of English, correcting and clarifying previously obtained estimates. Initially, Shannon (1951) estimated the entropy of printed English between 0.6 and 1.3 bits per character. Then Brown et al. (1992) gave an estimate of the upper bound of printed

texts in English equal to 1.75 bits per character, considering 128 ASCII characters. Next Teahan and Cleary (1996) estimated the average entropy of English texts at 1.46 bits per character. The entropy of the English language is 1.77 bits per character relative to Kontoyiannis (1997). Teahan and Cleary (1996) estimated the entropy of the English language from 0.94 to 1.72 bits per character, considering 32 characters of the alphabet. Calin (2020) estimated the entropy of modern English as 1.37 bits per character.

3 Corpus description

The corpus we analyze is based on text samples from the iWeb corpus of English language presented by Davies (2018) and contains about 100 million characters collected from web pages. Web corpora allow us to research many linguistic changes and reflections with minimal time lag. Unlike other large corpora from the web, the iWeb corpus was created in a systematic way and includes specific websites.

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

3.1 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 legal texts which excite in the English language are represented in the n-gram dictionaries.

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.

The number of n-grams extracted from the corpus is shown in Table 1.

Table 1. Corpus size.

n-grams	the number of extracted n-grams
10-grams	102 222 144
15-grams	102 222 139
20-grams	102 222 134
25-grams	102 222 129

The text corpus created and program tools are available on GitHub.[1]

[1] https://github.com/Nastasian/entropy/releases

4 Methodology

4.1 Dictionaries of n-grams

Within the n-gram model of the language, we generate the dictionaries. The dictionary is a set of n-grams arranged alphabetically, without repetition. We consider an n-gram as a sequence of n characters. The 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.

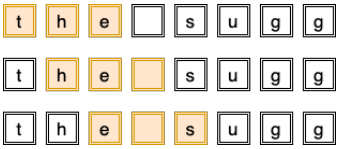

Fig. 1. The process of trigram selection with chaining.

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

The dictionary volume is the number of unique n-grams that remain after removing duplicates in the corpus. We fix the dictionary volumes and the number of n-grams that occur in the corpus only once for various n.

In this paper, we consider 4 n-gram model orders: 10-grams, 15-grams, 20-grams and 25-grams. We choose a step between the model orders of 5 symbols for more accurate extrapolation model construction. For low model orders, there are some statistics, as opposed to model orders over 10. It is quite difficult to count n-grams with an order higher than $25 - 30$. Since the study of the n-grams is conducted on a limited-size corpus, the coverage of n-grams decreases with increasing n.

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 legal n-grams that exist in the language and random n-grams that are impossible for the language.

4.2 Coverage and dictionary resizing

Coverage is the ratio of the volume of the constructed dictionary to the total number of different s-grams that exist in the language.

Technically, the coverage of the dictionary can be estimated by ratio of dictionary volume (number of n-grams in dictionary) and the number of all legal n-grams in the selected language. The problem with using this approach is that the exact number of all

legal n-grams of the language is unknown, especially for large n-grams. For an asymptotic estimate of the number of legal texts of fixed length, one could use the model of Shannon (1948), but the accuracy of this model is still not fully understood. In addition, such an approach would require relatively accurate estimates of the entropy of n-grams, and such results are still very few and the accuracy of the entropy estimate strongly depends on the type of text being evaluated. Thus, it is necessary to search for alternative methods for evaluating the coverage of dictionaries.

Since dictionaries are compiled on the corpus of limited length, their coverage is incomplete. This means that not all possible existing n-grams of the language are included in this dictionary. That is, type I errors are possible, when a legal 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, the n-gram dictionary is a tool for distinguishing between two statistical hypotheses:

- H_0 - n-gram is a legal text,
- H_1 - n-gram is a random sequence of characters.

The probability of the type I error to take a legal n-gram for a random set of characters is determined by the dictionary coverage τ:

$$\alpha = P(H_1|H_0) = 1 - \tau \tag{1}$$

The probability of the type II error to take a random set of characters for a legal n-gram is considered close to 0, since the proportion of forbidden n-grams that fall into the dictionary is negligibly small: $\beta = P(H_0|H_1) = 0$.

Therefore, it is necessary to evaluate the coverage of the dictionaries created, that is, to estimate what proportion of possible n-grams of the language our dictionaries cover.

In this study, we propose a theoretical approximate estimate of the coverage that is independent on empirical tests: $\tau = 1 - \frac{k_n}{K_n}$,

where K_n is the initial volume of n-gram dictionary, and k_n is the number of n-grams that occur in the corpus only once, τ is the theoretical coverage of n-gram dictionary.

This approach allows us to estimate approximately the amount of coverage of the average language corpus collected from a large number of independent texts. We assume that this coverage estimate is used to analyze a standard qualitative corpus with a large dictionary size. That is, the corpus collected from a large number of texts contains multiple repetitions of widely used and standard vocabulary. An exception may be highly specialized terms and vocabulary, which can be included in the corpus without repetition. Therefore, this approach is not suitable for the analysis of specialized and specific text corpora.

In addition, it is important to note that this approach to the assessment of coverage is developed for the practical analysis of the compiled corpuses and does not give a correct assessment of the coverage in all theoretically possible cases. For example, in degenerate cases, when the corpus a priori has a high coverage, but practically does not contain repetitions, or, conversely, is composed of multiple repetitions the same set of vocabulary, the considered approach gives an incorrect assessment of the coverage.

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_n units with k_n elements occurring once. Then $1-\frac{k_n}{K_n}$ is a fraction of repetitive elements in the dictionary, and $(1-\frac{k_n}{K_n})\cdot K_n$ 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, $(1-\frac{k_n}{K_n})\cdot K_n < K_n$. Based on the paper of Chase et al. (1994), it is known that up to a certain point, the number of new n-grams in the dictionary grows at a linear rate. To get a dictionary consisting of K repeated elements, we need to increase its volume by $\frac{1}{1-\frac{k_n}{K_n}}$ times.

Thus, a new dictionary with volume

$$\tilde{K}_n = \frac{K_n}{1-\frac{k_n}{K_n}} \tag{2}$$

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

Thus, $\tilde{K}_n$ is a theoretical estimate of volume for an n-gram dictionary which presumably covers most of all possible n-grams in a language.

4.3 Entropy of n-grams

We consider the theoretical estimate of the dictionary $\tilde{K}_n$ as some approximation of the number of all possible n-grams in the language. This means that we consider all out of dictionary 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 entropy of a random source may be calculated in a classical way of Shannon (1948) $\hat{H}_n = -\sum_{(a_{i_1},a_{i_2},\ldots,a_{i_n})} p(a_{i_1},a_{i_2},\ldots,a_{i_n})\cdot \log_2 p(a_{i_1},a_{i_2},\ldots,a_{i_n})$,

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 combinatorical approach for calculating entropy based on the dictionary volume. This idea is based on the Kolmogorov's combinatorial method and the Shannon's second theorem.

Let $M(n)$ be the number of all possible n-grams in a language with an alphabet of power A. Since the number of all distinct n-grams in this alphabet is estimated as $A^n = 2^{n\cdot\log_2 n} > M(n)$ is greater than the number of legal n-grams in the language, then there is a value H_n, such that $M(n) = 2^{n\cdot H_n}$, where $H_n < \log_2 A$, as presented in the paper of Shannon (1948).

With the growth of n, the value H_n tends to a certain limit. Let $n \to \infty$, then the value: H=$\lim_{n\to\infty} \frac{\log_2 M(n)}{n}$ is the language entropy. The existence of this limit is strictly proved in the framework of the stationary ergodic model of a random source.

The second theorem of Shannon (1948) gives an asymptotic estimate of the number of all possible n-grams: $M(n) = 2^{H \cdot n}$, where H is entropy of the language.

Let the dictionary volume $\tilde{K}_n$ be an approximation of the number of all possible n-grams in the language. Then the entropy of n-grams per character (bits/symbol) can be estimated as: $\mathrm{H}_n = \frac{\log_2 \tilde{K}_n}{n}$

For low orders of the n-gram model, while the value of H_n still decreases with the growth of n, the entropy value of H_n can be directly used to estimate the number of possible legal texts of length n symbols. Starting from some n, the value of H_n stabilizes and no longer changes with increasing n. To estimate the number of legal n-grams of this length, the value of the limit H_n is used. The methodology for finding the limit of the H_n value through extrapolation is described in the Section 5.2.

5 Results and discussion

5.1 Dictionary properties

In Figure 2, we present the results of the estimation of the n-gram dictionaries for short length texts. As said in Section 3.1, the shown values are the empirically obtained. In the diagram, we see the size of n-gram dictionaries depending on the volume of the source corpus.

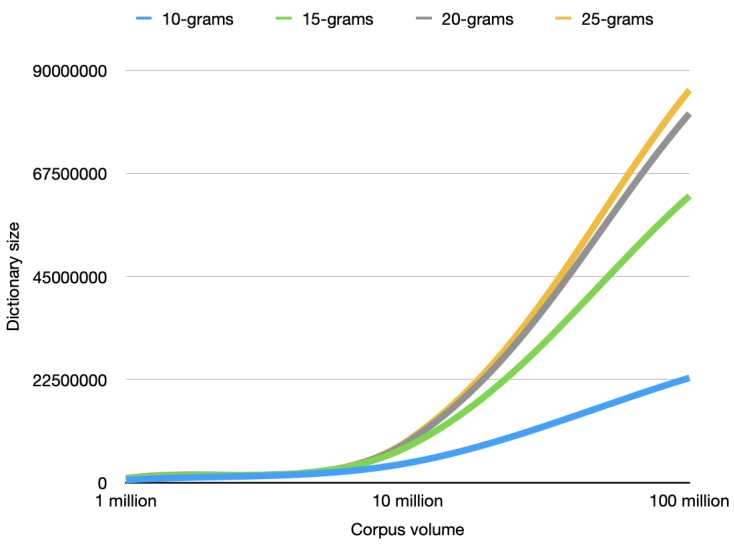

Fig. 2. Dictionaries of n-grams for shortlength texts.

Since the dictionary is a set of n-grams included in the text corpus, without taking into account duplicates, its size increases with the growth of n. The growth of the

dictionary with an increase in n is explained by the fact that the higher the order of the n-gram model, the less these n-grams are repeated in the original corpus. For high orders of n, there are already very few repetitions in the set of n-grams: many n-grams enter the corpus only once. For example, initially for 10-grams, a set which was extracted from a 100 million character corpus had a volume of 99999991 10-grams, for 25-grams - 99999976. However, after deleting the duplicates in each of the sets, there are respectively 22855480 10-grams and 85694340 25-grams in the dictionaries. The fact that the number of possible legal texts increases with its length is fully consistent with model of Shannon (1948) for estimating the number of legal texts in a language.

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 2.

Table 2. Initial coverage and new size of dictionary.

n-grams	Initial dictionary coverage, %	Theoretical dictionary volume
10-grams	51	22 millions
15-grams	32	149 millions
20-grams	21	386 millions
25-grams	16	606 millions

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 built an interpolation function. We have graphically represented the dependence of the dictionary size on the corpus size and noticed that it is similar to a square root function. Then we have constructed the interpolation function using *Wolfram Mathematica* and a non-linear fit form.

Listing 1.1. Interpolation construction

```
In[1]:= data={{10^6,389951},{10^7,2339589},{10^8,11252131}};
        nlm=NonlinearModelFit[data,a*Sqrt[x]+b,{a,b},x]
In[2]:=Show[ListPlot[data], Plot[nlm[x], {x, 10^6, 10^8}]]
```

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.

Equation 6 describes the interpolation function for 10-grams. $-1.14697 \cdot 10^6 + 1230.21\sqrt{x}$

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 example, 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.

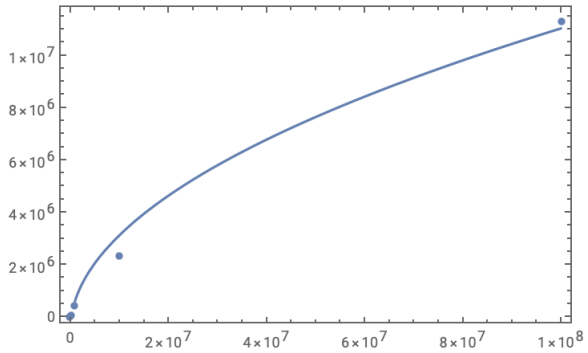

Fig. 3. 10-gram dictionary interpolation function.

5.2 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.

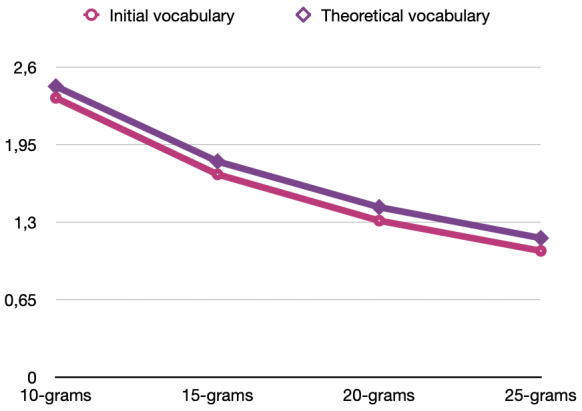

Fig. 4. Entropy per character for short length texts.

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: $\mathrm{F}_n - F_{n+1} = k \cdot (F_{n+1} - F_{n+2})$ with initial conditions $F_0 = H_{10}$, $F_1 = H_{15}$ and $F_2 = H_{20}$.

The coefficient k for the model is determined numerically in accordance with the experimentally obtained entropy values for segments of small length.

In this case, the value of k, which gives the best approximation, is $k \approx 0.62$. By increasing the value of n, a sequence of heuristic estimates of H_n is constructed, the experimental evaluation of which becomes difficult for a large length of a text segment. Starting from $n = 50$, the values of H_n are stabilized and no longer change with the length of the segment.

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

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 dictionary. Therefore, we can estimate the entropy H_n as the limit entropy for a long text.

5.3 Number of legal texts

For various studies, such as cipher systems, the number of possible legal texts of fixed length plays an important role.

The number of all distinct n-grams in the alphabet of power A is estimated as A^n. However, among this set, many n-grams are invalid for the selected language. As described in Section 4.3, the number of legal n-grams of high orders can be estimated as

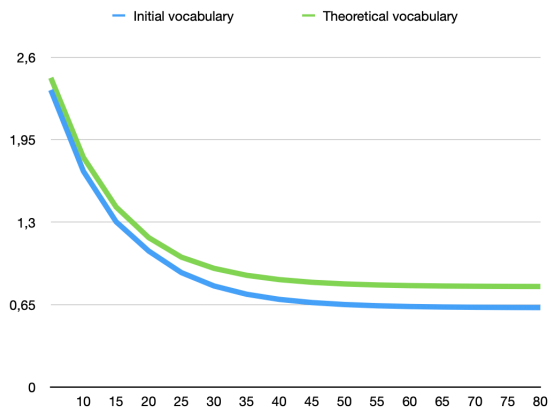

Fig. 5. Extrapolated entropy rate values.

$2^{H \cdot n}$, where H is the entropy of the language introduced in Section 4.3 and found in Section 5.2.

Using the entropy values obtained and the extrapolation model constructed, we have estimated the number of legal texts among all texts of fixed length. The results are shown in Table 3.

Therefore, the relative part of legal texts among all distinct texts of fixed length n can be estimated as: $2^{H \cdot n} \frac{}{A^n,}$ where H is the language entropy, A is the language alphabet, n is the text length.

Table 3. Number of legal texts.

n-gram model order	Number of legal texts	Relative part
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	$6.7 \cdot 10^{240}$	$2.7 \cdot 10^{-1222}$

The last finding is the proportion of legal n-grams among the total number of n-grams. In other words, it is the probability of finding a legal 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 legal 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.

6 Conclusion

In this paper, we have estimated the n-gram entropies of natural language texts and examined the number of legal possible 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 existing 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 legal long-length n-grams, so it is almost impossible to do empirically. The probability of finding a legal 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

Bellegarda, J. (2001). Robustness in statistical language modeling: Review and perspectives, *in* Junqua JC., v. N. G. (ed.), *Robustness in Language and Speech Technology. Text, Speech and Language Technology*, Vol. 17, Springer, Dordrecht, pp. 101–121.

Brown, P., Della Pietra, V., Della Pietra, S., Lai, J., Mercer, R. (1992). An estimate of an upper bound for the entropy of english, *Computational Linguistics* **18**(1), 31–40.

Calin, O. (2020). Statistics and machine learning experiments in english and romanian poetry, *Sci* **2**(4).

Chase, L., Rosenfeld, R., Ward, W. (1994). Error-responsive modifications to speech recognizers: Negative n-grams, *Third International Conference on Spoken Language Processing 1994*, International Speech Communication Association, Yokohama, Japan, pp. 827–830.

Chomsky, N. (1956). Three models for the description of language, *IRE Transactions on information theory* **2**(3), 113–124.

Cover, T., King, R. (1978). A convergent gambling estimate of the entropy of english, *IEEE Transactions on Information Theory* **24**(4), 413–421.

Davies, M. (2018). iweb: The 14 billion word web corpus, `https://www.english-corpora.org/iweb/`.

Florencio, D., Herley, C. (2007). A large-scale study of web password habits, *Proceedings of the 16th international conference on World Wide Web*, Association for Computing Machinery, New York, USA, pp. 657–666.

Guerrero, F. G. (2009). A new look at the classical entropy of written english, *arXiv preprint arXiv:0911.2284* .

Hahn, L. W., Sivley, R. M. (2011). Entropy, semantic relatedness and proximity, *Behavior research methods* **43**(3), 746–760.

Jaglom, A., Jaglom, I. (1973). *Probability and information. Processed and complemented*, Science (in Russ.), Moscow.

Jurafsky, D., Martin, J. (2009). *N-gram Language Models*, Pearson Prentice Hall, New Jersey.

Kolmogorov, A. (1993). Three approaches to the definition of the notion of amount of information, *in* Shiryayev, A. (ed.), *Selected Works of A. N. Kolmogorov. Mathematics and Its Applications (Soviet Series)*, Vol. 27, Springer, Dordrecht.

Kontoyiannis, I. (1997). The complexity and entropy of literary styles, *Technical report*, Department of Statistics, Stanford University.

Poncelas, A., Maillette de Buy Wenniger, G., Way, A. (2017). Applying n-gram alignment entropy to improve feature decay algorithms, *The Prague Bulletin of Mathematical Linguistics* **108**, 245–256.

Rosenfeld, R. (1995). Optimizing lexical and n-gram coverage via judicious use of linguistic data, *EUROSPEECH '95 Fourth European Conference on Speech Communication and Technology*, International Speech Communication Association, Madrid, Spain, pp. 1763–1766.

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell system technical journal* **27**(3), 379–423.

Shannon, C. E. (1951). Prediction and entropy of printed english, *Bell system technical journal* **30**(1), 50–64.

Teahan, W., Cleary, J. (1996). The entropy of english using ppm-based models, *Proceedings of Data Compression Conference - DCC '96*, IEEE, Snowbird, UT, USA, pp. 53–62.

White, H. (1967). Printed english compression by dictionary encoding, *Proceedings of the IEEE* **55**(3), 390–396.

Yadav, N., Joglekar, H., Rao, R. P., Vahia, M. N., Adhikari, R., Mahadevan, I. (2010). Statistical analysis of the indus script using n-grams, *PLoS One* **5**(3).

Received February 9, 2021 , revised July 21, 2021, accepted September 9, 2021