entropy 2021

sciforum

Abstract

Entropy Analysis of n-Grams and Estimation of the Number of Meaningful Language Texts †

Anastasia Malashina

HSE University, Moscow, Russia

† Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Published: 5 May 2021

When solving a number of information security problems, one of the problems is to estimate the number of possible meaningful texts of fixed length. To estimate this value, various approaches can be used, in each of which the key parameter is the information entropy. To estimate the number of short plaintexts, the entropy of n-grams is used. For long ones, in turn, we use the entropy of the language (specific entropy). n-grams, in this case, are n consecutive characters of meaningful text. The well-known information-theoretic approach allows us to obtain an asymptotic estimate of the meaningful text number based on the second Shannon theorem. In practice, to implement this approach, the text under study is presented in the form of a Markov source.

We consider a different approach to estimating the number of meaningful language texts, using the combinatorial method, the origins of which go back to the ideas of Kolmogorov. Representing a text as a set of independent n-grams, we experimentally estimate the number of semantic n-grams in a language by compiling dictionaries based on a large text corpus. In order to evaluate the I type errors of taking a meaningful n-gram for a random one, which inevitably occur during experimental evaluation, we developed a methodology for evaluating the coverage of the dictionary. We use this amount of coverage to refine and recalculate the original volume of the dictionary. Based on the number of meaningful n-grams of the language, we determine the entropy of short texts of various lengths. This sequence of estimates allows us to mathematically model the further change in the entropy function, extrapolate for long segments, and find the specific value of the entropy of the language.