Probability Analysis of the Vocabulary Size Dynamics Using Google Books Ngram Corpus *

Anastasia Pekina, Yulia Maslennikova, Vladimir Bochkarev

Kazan Federal University, Kazan, Russia
pekina.96@mail.ru

Abstract. The article introduces a method for determining a rate of appearance of new words in a language. The method is based on probabilistic estimates of the vocabulary size of a large text corpus. Backward predicted frequencies of rare words are estimated using linear models that are optimized by the maximum likelihood criteria. This approach provides more accurate estimations of frequencies for the earlier periods; the lower the frequency of the word during the analyzed period, the higher the benefit. A posteriori estimates of the frequency probability of appearance of new words were used to clarify the vocabulary size for different years and rate of appearance of new words. According to the proposed probabilistic model, it was shown that $> 30\%$ of investigated English and Russian word were appeared in the language before the moment when they were identified in the Google Books Ngram Corpus.

Keywords: Word usage frequencies · Prediction · Google Books Ngram.

1 Introduction

Despite the long history of studying languages, we still do not know, even approximately, how many words a specific language contains. Let's consider the English language. At present, the most complete published English dictionary, Oxford English Dictionary [1], comprises more than 600,000 words. However, it is obvious that it contains not all English words. For example, it does not contain extremely rare words (occurring less than 1 per billion words). Creation of Google Books Ngram, which contains more than 500 billion English words, brought hope to researchers to obtain a fairly complete list of words. In [2], an attempt was made to estimate the total number of words using this corpus. Estimations were obtained only at three points. According to their research, the language contained 554 thousand words in 1900, 597 thousand words in 1950 and 1022 thousand words in 2000. The article [2] shows graphs of the number of words obtained by linear extrapolation for the remaining years of the 20th century, without taking into account rare words. In [3], the authors evaluate the

* This work was supported by the Russian Foundation for Basic Research. Grant no. 17-29-09163 and the Russian Government Program of Competitive Growth of Kazan Federal University.

lexicon dynamics of the English language over the past 200 years, based on the ratio of the number of unique words that appeared in the core language to the number of all words for each decade using the Corpus of Historical American English and Google Books Ngram. The authors did not consider rare words that occur in the corpus less than 300 times over 10 years. In [4], much attention was paid to the analysis of the dynamics of the language core. The authors showed that since 1805 the actual English core has not significantly changed (words in the language core are being updated with a rate of about 30 words per year), and the speed of occurrence of words that do not enter the language core decreases with time. Thus, few works are devoted to the analysis of the active vocabulary in the early years (1800 and earlier), even fewer works take into account rare words that only enter into circulation. It should be noted that such words may exist in the language, but not appear in a certain year, due to the limited volume of texts in the given year.

In this paper, we propose a probabilistic model for clarifying the dynamics of word formation in English and Russian using the Google Books Ngram data. This probabilistic approach allows us to take into account the limited amount of texts related to the earlier period. The suggested method is based on prognostic estimations of the frequency of use of rare word forms in the past, which are then used to calculate more plausible estimations of the probability of occurrence of a word form in a lexicon, taking into account the size of the corpus.

2 Database

This research is based on the analysis of words usage dynamics from the Google books Ngram database. Frequencies of all unique 1-grams from the database were calculated. Then, English and Russian corpora were analyzed. The English corpus contained 5.3 million of unique words; the Russian corpus included 4.9 million of unique words. We analyzed dynamics of words that appeared in the corpora in 1800 year ($\sim$ 23,000 English words and $\sim$ 20, 000 Russian words).

The Google books Ngram dataset has been criticized for its reliance on inaccurate Optical character recognition, an overabundance of scientific literature, and for including large numbers of incorrectly dated and categorized texts [5]. Another issue is that Optical character recognition is not always reliable, and some characters may not be scanned correctly. In particular, systemic errors like the confusion of "s" and "f" can cause systemic bias. Although Google Ngram Viewer claims that the results are reliable from 1800 onwards [2].

There are few examples of the most popular incorrect 1-grams of the early English frequency dictionary: "a'dvance", "traditional1", "draw_", "knowtheir", "ossophagus" etc. 1-grams consist of numbers, not of the Latin letters, possible missing spaces and the replacement of letters. In this paper, the pre-processing of the investigated database was carried out. To check the early English 1-grams, the online Multitran dictionary was used, which is an Internet system containing electronic dictionaries of more than 14 languages and over 5 million terms in all language parts of the dictionary [6]. Each of the 23,000 selected English 1-grams

was checked in the Multitran dictionary. For the early Russian 1-grams, the dictionary "Open Corpora" was used [7]. This is the crowdsourcing project to create morphologically, syntactically and semantically marked corpus of texts in Russian, fully accessible to researchers. The corpus contains about 5.1 million words. Having checked the correctness, only 2161 of English 1-grams and 8452 of Russian 1-grams were selected for further analysis. A random checking of rare words removed from the investigated database showed that actually a large number of rare words that appeared in the corpora in 1800 were not correctly recognized or were misspelled in the original Google books Ngram database.

3 Methods and results

The proposed probabilistic approach is based on the idea of using the predicted estimations of 1-grams usage frequencies for refining them in early years. Typically, frequency estimates for later time are more reliable due to the larger volume of the database. On the left side of the Fig. 1 a graph of the number of books written in different years and included into the Google Books corpus is shown (English corpus version 2012.07.01). Therefore, based on the latest data, it is possible to predict the expected frequencies for earlier periods. This method is called "backward prediction".

There are many ways to estimate the prediction coefficients of the autoregression model; the particular way depends on the fluctuation distribution of the investigated time series. In our case, we are talking about estimating the usage frequencies of sufficiently rare words, and, consequently, we can expect that fluctuations are distributed according to the Poisson law, which is given by the probability function:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

The variance and mathematical expectation of a word usage frequency, distributed according to Poisson's law, are equal to the distribution parameter λ, that depends on the time, therefore we will use the symbol λ_t. The most accurate estimates of the parameter λ_t can be obtained using the maximum likelihood method (MLE) [6], when the log likelihood function of the following form is maximized:

$$\log L(\overrightarrow{\lambda_t}) = \log w(\overrightarrow{X}|\overrightarrow{\lambda_t}) = \sum_t X_t \log \lambda_t - \sum_t \lambda_t - \sum_t X_t! \tag{2}$$

Similar approach can be applied for a nonlinear prediction using artificial neural networks with maximum likelihood training. The approach is proposed in more detail in the paper [9].

Use of the maximum likelihood method significantly improves the accuracy of estimated frequencies for rare words if compared to the ordinary least squares procedure (MSE) [8]. A limiting form of the Poisson distribution is the Gaussian distribution. If the value of the parameter λ is large, then the results of MLE estimates will be similar to the results of the weighted MSE. Thus, for modelling

of the frequently used words, the MLE approach does not provide significant advantages. To compare the effectiveness of MLE and MSE methods, statistical modelling was carried out. We have artificially modeled the series of random numbers distributed according to Poisson's law, whose λ parameter changed with time according to the law $\lambda(t) \sim \exp(\alpha t)$. The values of the parameters $\lambda(t)$ were chosen by corresponding to real words frequencies from the base. After this, the parameters α were estimated using two methods. The results of calculations, as expected, show that the lower the frequency of word usage, the greater the gain from using the MLE method. For example, with an averaged frequency of 0.5 usage per year, the standard deviation of the estimate the parameter α is reduced 2 times compared to the usual estimate based on the average value of the empirical frequency.

The probability density function of calculated parameters for auto-regressive models for English 1-grams are shown on the right side of the Fig. 1. It can be seen that many 1-grams from pre-processed database (Density curve "AFTER") are widely used in later years, because the center of the probability density function corresponds to positive value, which is expected for the developing early language. The maximum of the density curve "BEFORE" (before the database pre-processing) is located in the negative range of values, since typos and words with recognition errors in the database do not tend to increase the usage frequencies (they are presented in the database with approximate constant small frequencies).

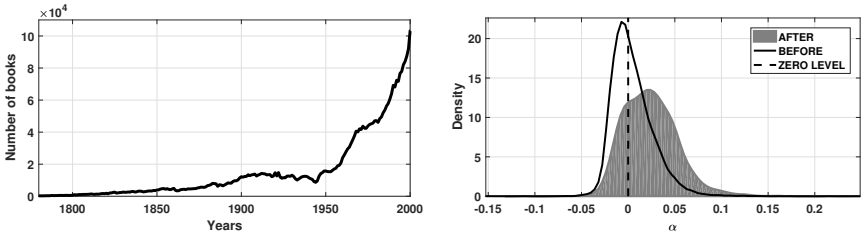

Fig. 1. The number of books written in different years and included into the Google Books corpus (on the left); probability density function of calculated parameters for autoregressive models for English 1-grams database before and after pre-processing (on the right)

Fig. 2 (on the left) shows the predicted usage frequencies of two rare English words 'shiftlessness' and 'tunnelling' using simple regression linear model of the first order with maximal likelihood optimization. The order of the model was chosen according to the size of the investigated time series (210 points for each words) because the data is very noisy and the use of higher-order models cannot be effective. This model leads to an exponential frequency dependence on time $\sim e^{\alpha}$. The dependence is plotted in logarithmic scale along the ordinate axis. Fig. 2 (on the right) shows the same plot for two rare Russian words. We should

note that these Russian and English words were rare in early years. In both cases, the backward predicted frequencies have positive value for the parameter of the exponential model, in other words, usage frequencies were increasing after 1800 year.

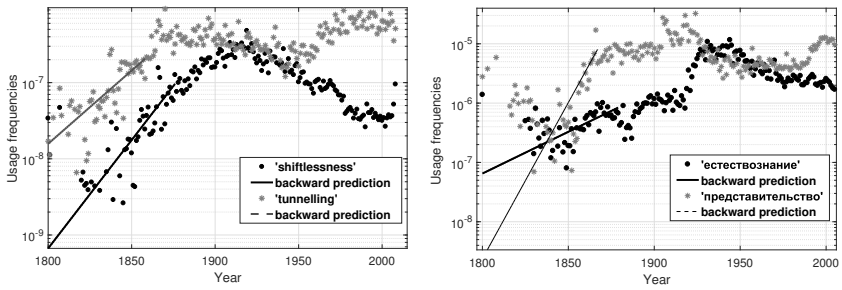

Fig. 2. Observed (markers) and backward predicted (solid lines) usage frequencies for two English (on the left) and Russian (on the right) words from the investigated database

After backward prediction of usage frequencies, it was shown that prediction errors are distributed approximately lognormally for both databases (English and Russian). Knowledge of backward predicted values and the error distribution law makes it possible to estimate the actual usage frequencies by the criterion of the maximum a posteriori probability. This criterion has a significant advantage over the estimates based on the mean value of empirical frequencies.

Updated information about usage frequencies for early years allows using this information to refine the actual volume of the lexicon and the speed of word formation. For example, we have registered a word in the corpus for the first time. The first possible reason that the word that was early used in the living language fell into the corpus because of the increase of its volume. The second reason is that it can be really a new word. By extrapolating usage frequencies to the previous years, we can calculate the probability that a word with such frequency could not be identified in a corpus of a known volume. Using such calculations for each word, allows us to specify the actual lexicon dynamic (and, correspondingly, the speed of appearance of new words) for different years.

Let $\hat{f}_t$ be the predicted in the past, the relative usage of a word for a certain year t, and N_t is the volume of the corpus for that year. Hence the probability that the word will not occur in this year in the case will be $(1 - \hat{f}_t)^{N_t}$. Since we will consider the frequency of word usage to be independent in different years, the total probability P_0 of the fact that the word did not appear in the body before 1800 will be written as a product of probabilities for different years:

$$P_0 = \prod_t (1 - \hat{f}_t)^{N_t} \tag{3}$$

The approach is presented in more detail in [10]. The null hypothesis is that the word had appeared in the language before the moment when it was first identified in the corpus. For example, the probability that the word 'shiftlessness' was not included in the database because of the small volume of the corpus before 1800 is 0.75, and this probability is 0.56 for the word 'tunnelling'. It can be seen that this probability is quite high.

The probability of 737 English words (34%) from the investigated database is > 0.5, in other words, these words were appeared before 1800 year with the probability > 0.5. Analyzing of predicted usage frequencies for these words, it was found that many of them (>74%) could appear before 1700 year, and only 26 % were born around 1800 year. $\sim$ 3,100 words (36%) from the Russian database show the probability > 0.5 (>70% of them could be born before 1700 year).

According to the proposed probabilistic model, it was shown that the date of the first appearance of a word in the corpus does not always coincide with the date of appearance of this word in the language. For example, more than 30% of words that first appeared in the corpus in 1800 were highly likely to have come into use much earlier, but did not enter the corpus before 1800 due to the insufficient size of the database.

References

1. OED Online Homepage, http://www.oed.com/. Last accessed 15 Apr 2018
2. Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., The Google Books Team, Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwang, J., Pinker, S., Nowak, M., Aiden, E.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science **311**(6014), 176–182 (2011)
3. Jatowt, A., Tanaka, K.:Large scale analysis of changes in english vocabulary over recent time. ACM International Conference Proceeding Series. 2523-2526. (2012)
4. Gerlach, M., Altmann, EG.: Stochastic Model for the Vocabulary Growth Natural Languages. Phys. Rev. X **3**(2), 021006 (2013)
5. Pechenick, E.: Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. PLoS ONE **10**(10), e0137041 (20151)
6. Multitran Homepage, https://www.multitran.ru/. Last accessed 15 Apr 2018
7. Open Corpora Homepage, http://opencorpora.org/. Last accessed 15 Apr 2018
8. Jackson, L.: Digital Filters and Signal Processing. 2nd edn. Kluwer Academic Publishers, Boston (1989)
9. Maslennikova, Y., Bochkarev, V., Voloskov, D.: Modelling of word usage frequency dy-namics using artificial neural network. Journal of Physics: Conference Series **490**(1), 012180 (2014)
10. Bochkarev, V.; Lerner, E.; Shevlyakova, A.: Deviations in the Zipf and Heaps laws in natural languages. Journal of Physics: Conference Series **490**(1), 012009 (2014)