2. Rosenfeld R. Optimizing lexical and n-gram coverage via judicious use of linguistic data // Proceedings of the Fourth European Conference on Speech Communication and Technology – Madrid, 1995.

3. Bellegarda J. R. Robustness in Statistical Language Modeling // Robustness in Language and Speech Technology , Springer Science+Business Media Dordrecht, 2001, pp. 104-106.

4. Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language // Conference on Intelligent Text Processing and Computational Linguistics – Mexico City, 2001.

УДК 519.722

Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов

А. Г. Малашина (Россия, г. Москва)
Национальный исследовательский университет «Высшая школа экономики»
e-mail: amalashina@hse.ru
А. Б. Лось (Россия, г. Москва)
Национальный исследовательский университет «Высшая школа экономики»
e-mail: alos@hse.ru

The construction and analysis of the Russian language models for a cryptographic algorithm research

A. G. Malashina (Russia, Moscow)
Higher School of Economics — National Research
e-mail: amalashina@hse.ru
A. B. Los (Russia, Moscow)
Higher School of Economics — National Research
e-mail: alos@hse.ru

1. Introduction

The word and n-gram distribution of natural language is studied in many areas: linguistics, game theory, and molecular biology. Corpus analysis of the language is also important in cryptographic protection of information, including the effectiveness of cryptographic algorithms. The procedures for recovery of discrete message parts are based on n-gram dictionaries of various lengths. In this regard, it is of particular importance to study their statistical properties, test the completeness and adequacy of the corpus used [1]. When studying the language corpus and compiling dictionaries, one of the main issues is the coverage of all possible text segments with the dictionary [2]. The coverage problem is significantly more complex for inflectional languages such as French and German, and especially Russian, compared to analytical languages such as English. Such languages require a larger vocabulary to achieve the required coverage [3]. The paper examines two language models of the Russian language: lexical and n-gram. In the lexical model of the language, the unit of analysis is tokens, that is, units of text, elements of separate writing. In the n-gram model, which is a special case of the lexical model, sequences of n characters or words are considered.

2. Language corpus

Corpus is a collection of texts used for language research via computer technologies. In this paper, a specialized corpus of the Russian language is created and reflects the narrow area of language usage. As the source material for compiling the corpus, news articles of recent years on political topics are used. These texts reflect the state of modern Russian, including the spoken language, that is, the corpus is syncronical.

After creating a text corpus, it is normalized, which consists of the following steps: 1) deleting html tags and reformatting them to *.txt; 2) recoding; 3) deleting all short forms of words excepting abbreviations; 4) deleting common names ; 5) filtering the text (deleting all characters except “а-я”, “.”, “,”,” “, cast to lowercase; 6) removing double spaces, repeating dots and commas, and spaces before dots and commas.

There is no metadata in the corpus, since it is not essential for further use of this corpus. The created corpus must meet two main criteria: completeness and representativeness. The completeness is determined by the coverage of this corpus. Optimization of coverage depends on the problems considered. First, the coverage depends on the text corpus volume that is used for compiling dictionaries, but this dependence becomes much less pronounced, so extrapolation by logarithmic functions is possible. 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 [3]. In General, the corpus is considered saturated when the sharp growth of new words stops with an increase in the corpus volume [2]. Representativeness is the corpus ability to adequately reflect the specifics of the narrow subject area [4]. Based on the collected news text corpus and in accordance with the considered language model, dictionaries are created and are subjected to statistical research.

3. The lexical model

Lexical models of the language are especially topical for speech recognition systems. The relevance of the maximizing coverage is due to the fact that each unknown word (also called out-of-vocabulary or OOV) creates another error in recognizing the current word. Moreover, each such error can generate a recognition error for the next word, creating a "ripple effect"of OOV words. Within the lexical model of the language, dictionaries are generated. The dictionary units are tokens, that is, units of text as elements of separate writing. The table below shows the size values of dictionaries that are created based on different size corpuses.

Corpus	Vocabulary volume	Empirical coverage	Theoretical coverage
10^4	836	0.45	20.1
10^5	5449	2.74	28.34
10^6	31895	13.78	39.68
10^7 with common names	163091	70.63	48.29
10^7 without common names	125162	64.45	44.98
10^8	*180000*	-	*59.28*
10^9	*230000*	-	*68.84*

Since the growth rate of dictionaries depending on the corpus size is close to the logarithmic function, the following extrapolation is performed:

$$21910.5 \cdot \ln(0.000034344 \cdot x). \tag{1}$$

Forecasted dictionary volume values are indicated in italics in the table. Graphically this relationship is shown in figure 1.

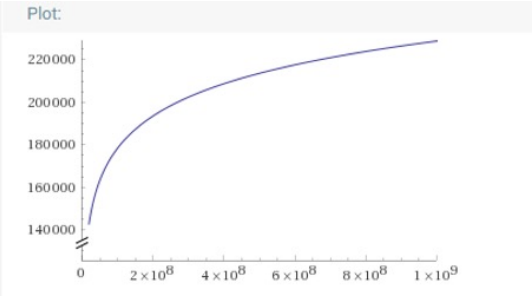

Рис. 1: The dictionary size depends on the corpus volume.

4. Zipf's Law

To test the quality and naturalness of the created Russian-language corpus (with a volume of 10^7 characters), compliance with the Zipf's law is checked. In accordance with this law, if all words in the corpus are ordered in descending order of their frequency, the frequency of word use will be inversely proportional to their ordinal rank [4]. The results are shown in figure 2.

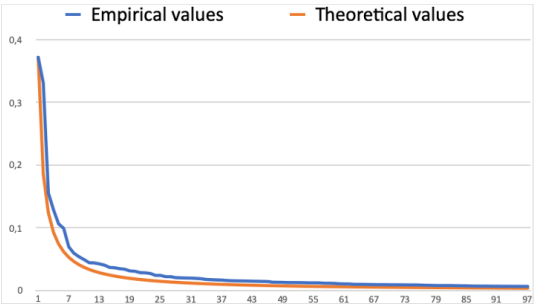

Рис. 2: Zipf's law.

As the graph clearly shows, the empirically obtained values deviate only slightly from the ideal values of Zipf's law. Therefore, we can conclude that the collected corpus generally satisfies this law and is quite natural and complete within the lexical model.

5. Lemmatization.

Lemmatization is the processing of a language corpus, as a result of which all the words included are reduced to a dictionary form [4]. Lemmatization is performed for the corpuses considered, and next the amount of coverage is estimated. Comparative results are shown in the table.

Corpus	Coverage (source)	Coverage (lemmatizated)
10^6	13.78	16.16
10^7 with common names	70.63	76.45
10^7 without common names	64.45	66.77

The results clearly demonstrate that the coverage values of lemmatized corpuses are higher than the coverage of the originals. This is due to the fact that lemmatization significantly reduces the inflection of the language, bringing it closer to the analytical one [3]. For example, the corpus considered contains the word "cats", while the test corpus contains only the word "cat". In other words, the same word is present in different corpuses in different forms. But from the point of view of automatic processing that accurately matches tokens, these are different words. This reduces the amount of coverage. Lemmatization eliminates this disadvantage and reduces the volume of the dictionary.

5. The N-gram model

In the n-gram model, the corpus (dictionary) units are sequences of n characters. Lexical models are a subset of n-gram language models.

N-gram coverage issues are relevant for speech recognition systems to maximize system performance. However, n-gram models are of particular importance for cryptography issues, since the boundaries between words in the cipher text are unknown, and attack by the lexical dictionary is impossible or extremely difficult. If an n-gram is missing from the dictionary, the language model may rely on lower-order n-grams, but they may not be appropriate for the current task. This is why recognition errors are much more common in the n-gram language model. Based on the corpuses collected, the coverage is evaluated, and the entropy of n-grams is calculated:

$$H_i = \frac{log_2 N_i}{i}. \tag{2}$$

where i is the length of an n-gram, and N_i is the number of n-grams in a word of length i characters. N-grams of 10, 15, 20, and 25 characters are considered. The results of the study are shown in the table and in figure 3.

Length of n-gramm	Dictionary volume		Entropy	
	10^6	10^7	10^6	10^7
10	795840	6217191	1.96	2.26
15	955193	9482897	1.32	1.55
20	983828	10372296	0.99	1.17
25	990430	10629589	0.80	0.93

The entropy values of n-grams are close to the real values for the Russian language. The coverage values are shown in the table below.

Length of n-gramm	Empirical coverage		Theoretical coverage	
	10^6	10^7	10^6	10^7
10	4.32	39.71	11.27	19.95
15	1.10	12.03	3.15	7.21
20	0.28	3.12	1.30	3.27
25	0.07	0.84	0.79	2.07

Therefore, the n-gram coverage is significantly lower than the lexical coverage. This confirms the fact that the study of the n-gram model is much more complex and optimizing the dictionary coverage requires an extra-large language corpus.

REFERENCES

1. Malashina A. G. The algorithm for recovering discrete message parts based on information about possible values of its characters. *Materialy konferencii. Mezhvuzovskaja nauchno-tehnicheskaja konferencija studentov, aspirantov i molodyh specialistov imeni E.V. Armenskogo* [Proc. Interuniversity scientific and technical conference of students, postgraduates and young specialists named after E. V. Armenian]. Moscow, 2019, pp. 215-217. (in Russian)

2. Rosenfeld R. Optimizing lexical and n-gram coverage via judicious use of linguistic data. *Proceedings of the Fourth European Conference on Speech Communication and Technology.* Madrid, 1995.

3. Bellegarda J. R. Robustness in Statistical Language Modeling. *Robustness in Language and Speech Technology, Springer Science+Business Media Dordrecht.* 2001, pp. 104-106.

4. Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language. *Conference on Intelligent Text Processing and Computational Linguistics.* Mexico City, 2001.

УДК 519.17

Обобщённые графы де Брейна

Ф. М. Малышев (Россия, г. Москва)
Математический институт им. В.А. Стеклова РАН
e-mail: malyshevfm@mi-ras.ru

Generalized de Bruijn graphs

F. M. Malyshev (Russia, Moscow)
Steklov Mathematical Institute of Russian Academy of Sciences
e-mail: malyshevfm@mi-ras.ru

1. Графы де Брейна [1], благодаря своему совершенству и широте их применения, занимают особое место среди специальных классов ориентированных графов. Рассматриваемые в настоящих тезисах обобщённые графы де Брейна сохраняют наиболее важные качества графов де Брейна. Под графом понимается ориентированный граф.

Граф де Брейна на $n = m^r$ вершинах, $n > 1$, $r \geqslant 1$, обозначаемый как $\partial B(m, r)$, является графом переходов состояний неавтономного автомата в виде регистра сдвига на r ячеек, в каждой ячейке которого может содержаться элемент конечного алфавита X, $|X| = m$. При входе $\varepsilon \in X$ состояние регистра сдвига $(x_1, \dots, x_r) \in S = X^r$ за один такт переходит в состояние $(\varepsilon, x_1, \dots, x_{r-1})$. Широкое распространение этот автомат получил, в частности, благодаря его свойству *быстрого обновления*, состоящего в том, что из любого начального состояния $\left(x_1^{(0)}, \dots, x_r^{(0)}\right) \in S$ за минимально возможное число тактов $r = \left\lceil \log_{|X|} |S| \right\rceil$ при независимых равновероятных входах $(\varepsilon_1, \dots, \varepsilon_r)$ автомат может оказаться в любом состоянии из S с одной и той же вероятностью $\frac{1}{|S|}$. Другие автоматы со свойством быстрого обновления можно строить как раз с помощью обобщённых графов де Брейна или, кратко, ∂-графов.

Определение 1. *Ориентированный граф* Γ *на* n *вершинах,* $n > 1$, *называется* ∂*-графом порядка* $r \geqslant 1$, *если для любых его вершин* v, w *существует единственный направленный путь длиной в* r *дуг из* v *в* w.

Наследование в ∂-графах свойств графов де Брейна, обусловленных определением 1, обеспечивает им широкие практические приложения. Автоматы с графами переходов внутренних состояний в виде ∂-графов могут быть использованы, наряду с регистрами сдвига, для генерации псевдослучайных последовательностей [2].

Обозначим через $Y = Y(\Gamma)$ матрицу смежности графа Γ. Ясно, что граф Γ на $n > 1$ вершинах является ∂-графом порядка $r \geqslant 1$ тогда и только тогда, когда $Y^r = J_n$, где J_n – матрица размера $n \times n$, все n^2 элементов которой равны 1. Имеется обширная литература (см., например, ссылки в [3]), относящаяся к сформулированной в [4] проблеме решения матричного уравнения $Y^r = \lambda J_n$, $\lambda \in \mathbb{Z}$, относительно целочисленных (0,1)-матриц Y. Много работ относится к решениям уравнения $Y^r = J_n$, являющимся циркулянтами.