СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2020. – № 1–2 (97).** – 113–126

УДК 519.6 DOI: 10.17212/2307-6879-2020-1-2-113-126

ФОРМИРОВАНИЕ ПСЕВДОСЛУЧАЙНЫХ ТЕКСТОВ НА ОСНОВЕ ЧАСТОТНЫХ ХАРАКТЕРИСТИК ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА*

Ю.А. КОТОВ[1], О.В. САНИНА[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат физико-математических наук, доцент, доцент кафедры защиты информации. E-mail: kotov@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: lyalysa@gmail.com*

В статье обсуждается вопрос генерации псевдослучайных текстов на основе частотных характеристик текстов естественного языка. Для генерации рассмотрены частотные характеристики и их значения для текстов на английском и русском языках: распределение униграмм и биграмм по частоте появления в тексте, распределение слов по длине. Предложен алгоритм генерации псевдослучайных текстов на основе данных частотных характеристик. Дана экспериментальная оценка сгенерированных текстов по алгоритму идентификации языка текста.

Ключевые слова: псевдослучайный текст, униграмма, биграмма, распределение Пуассона

ВВЕДЕНИЕ

Генерация текстов, в том числе автоматическая, сегодня применяется во многих областях: для формирования ответов на запрос пользователя [1, 2], автоматизированного написания новостных статей в журналистике [3, 4], аутентификации пользователя по характеристикам голоса [5]. Тексты, сгенерированные для решения таких задач, обладают семантикой и в общем случае оцениваются именно по ней. Отдельного внимания заслуживает генерация так называемых случайных текстов. Такие тексты могут быть сгенерированы для проверки шрифтов или макета (например, известный текст «Lorem ipsum» [6]). Известны случаи написания «псевдотекстов» для выявления журналов с недобросовестным рецензированием [7, 8].

* Статья получена 25 мая 2020 г.

Среди современных подходов к генерации случайных текстов широко используются нейронные сети [9]. Однако такие подходы требуют наличия обучающей выборки и вычислительно затратны.

Под псевдослучайным текстом в работе (псевдотекстом, случайным текстом) будем понимать текст, составленный по частотным характеристикам текстов естественного языка, но не имеющий семантической значимости и в общем случае не соответствующий правилам естественного языка. Предложенные псевдослучайные тексты могут найти применение в ряде задач для проверки точности оптического распознавания символов (OCR) [10], создания речеподобных помех, защиты речевой информации [11] и т. д.

Решение задачи по генерации псевдослучайных текстов, таким образом, сводится к решению задачи по генерации псевдослучайных последовательностей чисел.

1. ПОСТАНОВКА ЗАДАЧИ

Целью работы является определение набора частотных характеристик и составление алгоритма для генерации псевдослучайных текстовых последовательностей, таких что при использовании основных частотных характеристик данные последовательности распознаются как тексты на естественном языке.

Для решения поставленной задачи необходимо:

- определить набор частотных характеристик, эталонное распределение которых будет применяться для генерации случайных текстов;
- сгенерировать случайные тексты с частотными характеристиками текстов русского и английского языка и оценить их путем анализа на принадлежность к русскому или английскому языку на основе ряда частотных характеристик.

Для генерации псевдослучайных текстов будем использовать следующие частотные характеристики: распределение униграмм и биграмм по частоте встречаемости, распределение словоформ словника текстов по длине.

1.1. ЧАСТОТА ВСТРЕЧАЕМОСТИ УНИГРАММ

Самым первым и простым способом имитации текстов естественного языка является распределение униграмм текста по частоте встречаемости. В табл. 1 представлено распределение униграмм русскоязычных (мощность алфавита $N_A = 31$, «Ё» = «Е», «Ъ» = «Ь») и англоязычных ($N_A = 26$) текстов

объемом 5000 знаков по количеству появлений в тексте знаков C_i [12]. Аналогичные характеристики могут быть получены для текстов другого объема.

Таблица 1

Распределение униграмм по частоте встречаемости в русскоязычных и англоязычных текстах объемом 5000 знаков

Русский язык						Английский язык					
Буква	C_i	Буква	C_i	Буква	C_i	Буква	C_i	Буква	C_i	Буква	C_i
« »	670	Д	136	Ч	47	« »	733	C	149	Z	7
О	472	М	131	Ь	63	E	574	H	144	Q	5
Е	378	П	121	Х	53	T	393	P	116	J	4
А	325	У	110	Ч	47	R	371	M	106	–	–
И	325	Ы	94	Ш	37	I	362	F	83	–	–
Н	278	З	84	Ж	32	O	292	W	73	–	–
Т	278	Я	84	Ю	32	A	278	B	60	–	–
С	236	Г	74	Ц	21	S	272	G	57	–	–
В	210	Б	74	Щ	16	N	271	V	48	–	–
Р	199	Й	68	Э	11	D	193	Y	44	–	–
Л	184	Ь	63	Ф	11	L	187	X	15	–	–
К	147	Х	53	–	–	U	151	K	12	–	–

Несмотря на то что значения индекса совпадения в текстах, составленных по униграммному распределению, удовлетворяют граничным условиям [13], такие тексты всё еще могут быть отклонены как тексты на естественном языке на основе количества используемых биграмм, поскольку в текстах естественного языка используется ограниченное число биграмм (в частности, отсутствуют так называемые «запрещенные биграммы»).

1.2. ЧАСТОТА ВСТРЕЧАЕМОСТИ БИГРАММ

В [14] было показано, что в русскоязычных текстах даже объемом 350 тысяч знаков используется около 88 % всех возможных биграмм текстов. В связи с этим необходимо получить распределение биграмм в текстах естественного языка для дальнейшего использования в генерации псевдослучайных текстов. Таблицы частоты встречаемости и сочетаемости биграмм для русского и английского языка приведены в литературе, например в [15]. Однако даже такого набора частотных характеристик недостаточно, поскольку слова в таких текстах не согласуются с распределением Пуассона и, следовательно, могут быть распознаны по соответствующим критериям [16], связанным с длиной слов, используемых в текстах на естественном языке.

1.3. РАСПРЕДЕЛЕНИЕ СЛОВ ПО ДЛИНЕ

Известно, что распределение слов словника текста по длине согласуется с распределением Пуассона (рис. 1) [16, 17]. Тогда в тексте, в котором распределение частотных характеристик приближено к распределению частотных характеристик текста на естественном языке, должно учитываться данное распределение.

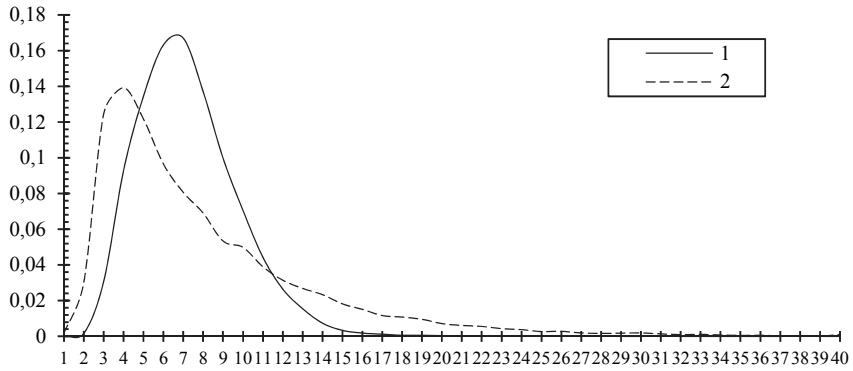

Рис. 1. Распределение словоформ словника:

1 – текста английского языка по длине; *2* – псевдослучайного текста, не учитывающего распределения слов

Таблица 2

Распределение слов по длине в текстах объемом 5000 знаков на русском и английском языке

Русский язык								Английский язык							
l	N_l	l	N_l	l	N_l	l	N_l	l	N_l	l	N_l	l	N_l	l	N_l
1	103	11	28	21	0	31	0	1	3	11	6	21	1	31	0
2	89	12	16	22	0	32	0	2	18	12	3	22	0	32	0
3	93	13	10	23	1	33	0	3	60	13	2	23	0	33	0
4	78	14	7	24	1	34	0	4	98	14	1	24	0	34	0
5	99	15	5	25	0	35	0	5	71	15	1	25	0	35	0
6	119	16	2	26	0	36	0	6	66	16	0	26	0	36	0
7	74	17	0	27	0	37	0	7	53	17	0	27	0	37	0
8	63	18	3	28	0	38	0	8	33	18	0	28	0	38	0
9	68	19	0	29	0	39	0	9	23	19	0	29	0	39	0
10	37	20	0	30	0	40	0	10	11	20	0	30	0	40	0

Распределение уникальных словоформ словника англоязычного и русскоязычного текстов объемом 5000 знаков по длине приведено в табл. 2, где l – длина словоформ, N_l – количество словоформ длины l.

Для генерации текстов можно также учитывать распределение Ципфа [18, 19] и другие закономерности языка.

2. АЛГОРИТМ ГЕНЕРАЦИИ ПСЕВДОСЛУЧАЙНЫХ ТЕКСТОВ

Алгоритм генерации псевдослучайных текстов на основе рассмотренных частотных характеристик текстов естественного языка (количество появлений униграмм, биграмм и слов каждой длины) реализуется в несколько этапов:

1) формируются эталоны распределений выборочных частотных характеристик;

2) случайным образом определяется первая буква в генерируемом псевдослучайном тексте;

3) на основе таблицы биграмм случайным образом определяется следующая буква в зависимости от предыдущей;

4) на основе распределения слов по длине случайным образом определяется длина каждого слова, в зависимости от которой определяется место пробела в тексте.

Блок-схема предложенного алгоритма представлена на рис. 2.

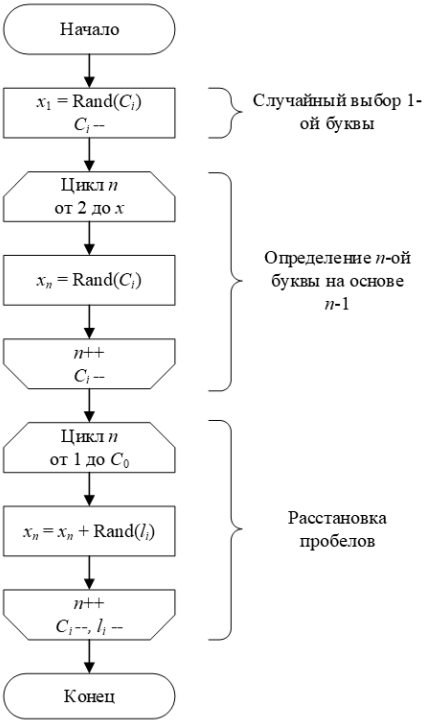

Рис. 2. Блок-схема алгоритма генерации псевдослучайных текстов

Для оценки полученных таким образом псевдослучайных текстов будем использовать алгоритм идентификации русскоязычных текстов, предложенный в [13].

По указанному алгоритму были сгенерированы выборки псевдослучайных текстов русского и английского языка. Тексты в выборках менялись в зависимости от объема. Характеристика составленных выборок приведена в табл. 3,

где K_1 – количество полученных псевдослучайных текстов русского языка объемом x знаков, K_2 – количество псевдослучайных текстов английского языка объемом x знаков.

Таблица 3

Характеристика выборок псевдослучайных текстов на русском и английском языке

x	K_1	K_2	x	K_1	K_2	x	K_1	K_2
Группа 1			Группа 2			90 000	100	100
200	100	100	2000	100	100	110 000	100	100
400	100	100	4000	100	100	Всего: 3	600	600
600	100	100	6000	100	100	Группа 4		
800	100	100	8000	100	100	100 000	30	30
1000	100	100	10 000	100	100	150 000	30	30
1200	100	100	Всего: 2	500	500	200 000	30	30
1400	100	100	Группа 3			250 000	30	30
1600	100	100	10 000	100	100	300 000	30	30
1800	100	100	30 000	100	100	350 000	30	30
2000	100	100	50 000	100	100	Всего: 4	180	180
Всего: 1	1000	1000	70 000	100	100	Итого:	2280	2280

2.1. ПРОВЕРКА ПСЕВДОСЛУЧАЙНЫХ ТЕКСТОВ

Для определения принадлежности текста к языку были использованы следующие частотные характеристики: количество используемых букв и биграмм, индекс совпадений.

Сгенерированные русскоязычные псевдослучайные тексты (см. табл. 3) были проверены на алгоритме определения принадлежности текста к русскому языку по данным характеристикам [13]. При проверке использовались следующие критерии.

Интервалы по количеству используемых букв алфавита в русскоязычных текстах (1):

$$
\begin{aligned}
&1)\ 200 \le x, && 21 \le Z(x) \le 32;\\
&2)\ 200 < x \le 600, && 25 \le Z(x) \le 32;\\
&3)\ 600 < x < 1000, && 27 \le Z(x) \le 32;\\
&4)\ 1000 < x \le 1400, && 28 \le Z(x) \le 32;\\
&5)\ 1400 < x \le 2000, && 29 \le Z(x) \le 32;\\
&6)\ 2000 < x \le 4000, && 30 \le Z(x) \le 32;\\
&7)\ 4000 < x, && 31 \le Z(x) \le 32.
\end{aligned}
\qquad (1)
$$

Количество биграмм в русскоязычных текстах не должно превышать граничное в соответствующей точке шкалы [13]:

$$y(x)=\begin{cases}3.16\cdot10^{-4}x+0.09, & 200\le x<800;\\ 1.30\cdot10^{-4}x+0.23, & 800\le x<2000;\\ 5.70\cdot10^{-5}x+0.38, & 2000\le x<4000;\\ 2.19\cdot10^{-5}x+0.52, & 4000\le x<10\ 000;\\ 4.86\cdot10^{-6}x+0.70, & 10\ 000\le x<30\ 000;\\ 3.01\cdot10^{-7}x+0.83, & 30\ 000\le x\le 350\ 000.\end{cases} \quad (2)$$

Граничные значения индекса совпадения для русскоязычных текстов находятся в интервалах [13]:

$$\begin{aligned}&1)\ 200\le x<800, && 0.0495<I_c(x)<0.0683;\\ &2)\ 800\le x<8000, && 0.05187<I_c(x)<0.06587;\\ &3)\ 8000\le x<100\ 000, && 0.05415<I_c(x)<0.06175;\\ &4)\ 100\ 000\le x\le 350\ 000, && 0.05481<I_c(x)<0.06069.\end{aligned} \quad (3)$$

Русскоязычные тексты (см. табл. 3), сгенерированные указанным образом, распознавались как русскоязычные в 100 % случаев, начиная с объема x = 1800 знаков. Тексты объемом от 1600 до 1800 знаков распознавались как русскоязычные только в 70 % случаев. В текстах объемом от 1400 до 1600 знаков 6 % текстов распознавались как русскоязычные. В текстах меньшего объема все тексты отбрасывались как нерусскоязычные.

Аналогично русскоязычным псевдослучайным текстам был проведен эксперимент по определению принадлежности текста к английскому языку.

Интервалы по количеству используемых букв алфавита англоязычных текстов:

$$\begin{aligned}&1)\ 200\le x, && 18\le Z(x)\le 27;\\ &2)\ 200<x\le 400, && 20\le Z(x)\le 27;\\ &3)\ 400<x<1000, && 21\le Z(x)\le 27;\\ &4)\ 1000<x\le 1800, && 22\le Z(x)\le 27;\\ &5)\ 1800<x\le 4000, && 23\le Z(x)\le 27;\\ &6)\ 4000<x\le 10\ 000, && 24\le Z(x)\le 27;\\ &7)\ 10\ 000<x, && 26\le Z(x)\le 27.\end{aligned} \quad (4)$$

В качестве граничных значений для английского языка по количеству биграмм использована прямая:

$$y(x)=\begin{cases}3.39\cdot10^{-4}x+0.12, & 200\le x<800;\\ 1.25\cdot10^{-4}x+0.29, & 800\le x<2000;\\ 3.74\cdot10^{-5}x+0.46 & 2000\le x<4000;\\ 2.00\cdot10^{-5}x+0.53, & 4000\le x<10000;\\ 5.18\cdot10^{-6}x+0.68, & 10\ 000\le x<30\ 000;\\ 4.02\cdot10^{-7}x+0.83, & 30\ 000\le x\le 350\ 000.\end{cases} \quad (5)$$

Граничные значения индекса совпадения:

$$\begin{array}{ll}1)\ 200\le x<600, & 0.0541<I_c(x)<0.0778;\\ 2)\ 600\le x<1400, & 0.0584<I_c(x)<0.0738;\\ 3)\ 1400\le x<30\ 000, & 0.0592<I_c(x)<0.0729;\\ 4)\ 30\ 000\le x\le 350\ 000, & 0.0613<I_c(x)<0.0700.\end{array} \quad (6)$$

Англоязычные тексты, сгенерированные по указанным характеристикам, распознавались как англоязычные в 100 % случаев, начиная с объема x = 1600 знаков. Тексты объемом от 1400 до 1600 знаков распознавались как англоязычные только в 65 % случаев. В текстах меньшего объема все тексты распознавались как неанглоязычные.

Кроме того, были проанализированы тексты, в которых генерация производилась исключительно по частоте встречаемости знаков. Для этого были использованы выборки с такими текстами, характеристика которых указана в табл. 4, где K_1 – количество псевдослучайных текстов русского языка объемом x знаков, K_2 – количество псевдослучайных текстов английского языка.

Таблица 4

Характеристика выборок псевдослучайных текстов, сгенерированных на основе частоты встречаемости знаков

x	K_1	K_2	x	K_1	K_2	x	K_1	K_2
Группа 1			Группа 2			90 000	50	100
200	100	100	2000	100	100	110 000	50	100
400	100	100	4000	100	100	Всего: 3	300	600

Окончание табл. 4

x	K_1	K_2	x	K_1	K_2	x	K_1	K_2
600	100	100	6000	100	100	Группа 4		
800	100	100	8000	100	100	100 000	8	30
1000	100	100	10 000	100	100	150 000	8	30
1200	200	100	Всего: 2	500	500	200 000	8	30
1400	100	100	Группа 3			250 000	8	30
1600	100	100	10 000	50	100	300 000	8	30
1800	100	100	30 000	50	100	350 000	8	30
2000	–	100	50 000	50	100	Всего: 4	48	180
Всего: 1	1000	1000	70 000	50	100	Итого:	1848	2280

В текстах выборки (табл. 4) объемом $x \leq 200$ знаков 94 % текстов были распознаны как русскоязычные. В текстах большего объема ($x > 200$) 100 % текстов распознаны как нерусскоязычные по количеству используемых биграмм. Таким образом, применение такого способа генерации возможно только для текстов малого объема ($x \leq 200$), в остальных случаях определить, что данные тексты являются псевдослучайными, можно на основеи числа используемых биграмм.

Аналогично русскоязычным текстам, в англоязычных текстах (табл. 4) также было проведено распознавание языка. Только 12 % от общего числа текстов объемом $x \leq 200$ знаков были распознаны как англоязычные, остальные 88 % отклонены по количеству используемых биграмм. В текстах большего объема ($x > 200$) все тексты распознаны как неанглоязычные по количеству используемых биграмм.

ЗАКЛЮЧЕНИЕ

Предложенный алгоритм генерирует тексты, 100 % которых, начиная с объема в 2000 знаков, распознаются как тексты на естественном языке. Полученный в работе алгоритм позволяет быстро генерировать псевдослучайные тексты с частотными характеристиками текстов на естественном языке.

Предложенный способ формирования псевдослучайных текстов может найти применение в проверке точности оптического распознавания символов, создании речеподобных помех для защиты речевой информации и в ряде других задач по формальному анализу текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Building a chatbot with serverless computing / M. Yan, P. Castro, P. Cheng, V. Ishakian // Proceedings of the 1st International Workshop on Mashups of Things and APIs. – Trento, Italy, 2016. – P. 1–4. – DOI: 10.1145/3007203.3007217.

2. A new chatbot for customer service on social media / A. Xu, Zh. Liu, Yu. Guo, V. Sinha, R. Akkiraju // 2017 CHI Conference on Human Factors in Computing Systems. – Denver, Colorado, USA, 2017. – P. 3506–3510. – DOI: 10.1145/3025453.3025496.

3. *Lokot T., Diakopoulos N.* News bots: automating news and information dissemination on Twitter // Digital Journalism. – 2016. – Vol. 4, N 6. – P. 682–699. – DOI: 10.1080/21670811.2015.1081822.

4. *Carlson M.* The robotic reporter: automated journalism and the redefinition of labor, compositional forms, and journalistic authority // Digital Journalism. – 2015. – Vol. 3, N 3. – P. 416–431. – DOI: 10.1080/21670811.2014.976412.

5. *Cheng F., Wang Sh.-L., Liew A.W.-Ch.* Visual speaker authentication with random prompt texts by a dual-task CNN framework // Pattern Recognition. – 2018. – Vol. 83. – P. 340–352. – DOI: 10.1016/j.patcog.2018.06.005.

6. Description of the "Lorem ipsum dolor sit amet" text that appears in Word Help. – URL: https://support.microsoft.com/en-gb/help/114222/description-of-the-lorem-ipsum-dolor-sit-amet-text-that-appears-in-wor (accessed: 08.07.2020).

7. *Wilson S.* Towards the simulation of sensor networks [authenticity and untruthful practice] // The Official: International Journal of Contemporary Humanities. – 2017. – Vol. 2, N 1. – P. 1–15.

8. *Melchior J.K.* Fake news comes to Academia: how three scholars gulled academic journals to publish hoax papers on ‘grievance studies’. – URL: https://www.wsj.com/articles/fake-news-comes-to-academia-1538520950 (accessed: 08.07.2020).

9. Thematic texts generation issues based on recurrent neural networks and word2vec / V. Fomenko, H. Loutskii, P. Rehida, A. Volokyta // Technical Sciences and Technologies. – 2017. – N 4. – P. 110–115. – DOI: 10.25140/2411-5363-2017-4(10)-110-115.

10. *Lopresti D.* Optical character recognition errors and their effects on natural language processing // International Journal on Document Analysis and Recognition (IJDAR). – 2009. – Vol. 12, N 3. – P. 141–151. – DOI: 10.1007/s10032-009-0094-8.

11. Создание речеподобной помехи на основе связных текстов / В.А. Трушин, Д.Е. Попов, М.А. Кунгуров, Д.Л. Марченко // Проблемы правовой и технической защиты информации. – 2018. – № 6. – С. 79–84.

12. *Абденов А.Ж., Котов Ю.А., Санина О.В.* Значения некоторых униграммных характеристик русскоязычных текстов // Научный вестник НГТУ. – 2017. – № 2. – С. 146–162. – DOI: 10.17212/1814-1196-2017-2-146-162.

13. *Kotov Yu.A., Sanina O.V.* Criteria and algorithm for the Russian language text recognition based on the frequency characteristics set // XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE 2018). – Novosibirsk, 2018. – P. 175–179. – DOI: 10.1109/APEIE.2018.8545877.

14. *Котов Ю.А., Санина О.В.* Значения некоторых биграммных характеристик русскоязычных текстов // Вестник СибГУТИ. – 2017. – № 4. – С. 24–34.

15. *Жданов О.Н., Куденкова И.А.* Криптоанализ классических шифров: лабораторный практикум / Сибирский государственный аэрокосмический университет им. М.Ф. Решетнева. – Красноярск, 2008. – 107 с.

16. *Котов Ю.А., Санина О.В.* Идентификация пробела при неизвестной знаковой кодировке в русскоязычных текстах // Вестник СибГУТИ. – 2018. – № 4. – С. 48–60.

17. *Smith R.D.* Distinct word length frequencies: distributions and symbol entropies // Glottometrics: Studies in Quantítative Linguistics. – 2012. – N 23. – P. 7–22.

18. *Moreno-Sanchez I., Font-Clos F., Corral A.* Large-scale analysis of Zipf's law in English texts // PLoS One. – 2016. – Vol. 11, N 1. – DOI: 10.1371/journal.pone.0147073.

19. *Ferrer-i-Cancho R., Elvevag B.* Random texts do not exhibit the real Zipf's law-like rank distribution // PLoS One. – 2010. – Vol. 5, N 3. – DOI: 10.1371/journal.pone.0009411.

Котов Юрий Алексеевич, кандидат физико-математических наук, доцент, доцент кафедры защиты информации Новосибирского государственного технического университета. Основное направление научных исследований – информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Имеет более 35 публикаций. E-mail: kotov@corp.nstu.ru

Санина Ольга Валерьевна, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Основное направление научных исследований – криптография на основе теории сложности, в частности безопасность протоколов обмена ключами. Имеет более десяти научных публикаций. E-mail: lyalysa@gmail.com

DOI: 10.17212/2307-6879-2020-1-2-113-126

Generating pseudo-random texts based on the frequency characteristics of texts in natural languages*

Yu.A. Kotov[1], O.V. Sanina[2]

[1] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, candidate of physical and mathematical sciences, docent, docent of the information security department. E-mail: kotov@corp.nstu.ru*
[2] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, Novosibirsk, 630073, Russian Federation, master's student of the computer engineering department. E-mail: lyalysa@gmail.com*

The paper discusses generation of pseudo-random texts based on frequency characteristics of texts in natural languages. The follow frequency characteristics of texts and their values for the Russian and English languages are considered for generation: the distribution of unigrams and bigrams over frequency of occurrence in texts, the distribution of words over the length. Based on the considered frequency characteristics, an algorithm for generating pseudo-random texts is suggested. Texts generated according to the algorithm are studied in experiments of language recognition in texts.

Keywords: pseudo-random texts, unigram, bigram, Poisson distribution

REFERENCES

1. Yan M., Castro M., Cheng M., Ishakian V. Building a chatbot with serverless computing. *Proceedings of the 1st International Workshop on Mashups of Things and APIs*, Trento, Italy, December 2016, pp. 1–4. DOI: 10.1145/3007203.3007217.

2. Xu A., Liu Zh., Guo Yu., Sinha V., Akkiraju R. A new chatbot for customer service on social media. *2017 CHI Conference on Human Factors in Computing Systems*, Denver, Colorado, USA, 2017, pp. 3506–3510. DOI: 10.1145/3025453.3025496.

3. Lokot T., Diakopoulos N. News bots: automating news and information dissemination on Twitter. *Digital Journalism*, 2016, vol. 4, no. 6, pp. 682–699. DOI: 10.1080/21670811.2015.1081822.

4. Carlson M. The robotic reporter: automated journalism and the redefini-tion of labor, compositional forms, and journalistic authority. *Digital Journalism*, 2015, vol. 3, no. 3, pp. 416–431. DOI: 10.1080/21670811.2014.976412.

5. Cheng F., Wang Sh.-L., Liew A.W.-Ch. Visual speaker authentication with random prompt texts by a dual-task CNN framework. *Pattern Recognition*, 2018, vol. 83, pp. 340–352. DOI: 10.1016/j.patcog.2018.06.005.

* *Received 25 May 2020.*

6. *Description of the "Lorem ipsum dolor sit amet" text that appears in Word Help*. Available at: https://support.microsoft.com/en-gb/help/114222/description-of-the-lorem-ipsum-dolor-sit-amet-text-that-appears-in-wor (accessed 08.07.2020).

7. Wilson S. Towards the simulation of sensor networks [authenticity and untruthful practice]. *The Official: International Journal of Contemporary Humanities*, 2017, vol. 2, no. 1, pp. 1–15.

8. Melchior J.K. *Fake news comes to Academia: how three scholars gulled academic journals to publish hoax papers on 'grievance studies'*. Available at: https://www.wsj.com/articles/fake-news-comes-to-academia-1538520950 (accessed 08.07.2020).

9. Fomenko V., Loutskii H., Rehida H., Volokyta H. Thematic texts generation issues based on recurrent neural networks and word2vec. *Technical Sciences and Technologies*, 2017, no. 4, pp. 110–115. DOI: 10.25140/2411-5363-2017-4(10)-110-115.

10. Lopresti D. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 2009, vol. 12, no. 3, pp. 141–151. DOI: 10.1007/s10032-009-0094-8.

11. Trushin V.A., Popov D.E., Kungurov M.A., Marchenko D.L. Sozdanie rechepodobnoi pomekhi na osnove svyaznykh tekstov [Generation of speech-like interfere based on coherent texts]. *Problemy pravovoi i tekhnicheskoi zashchity informatsii = Problems of law and technical information security*, 2018, no. 6, pp. 79–84.

12. Abdenov A.Zh., Kotov Yu.A., Sanina O.V. Znacheniya nekotorykh unigrammnykh kharakteristik russkoyazychnykh tekstov [Values of some unigram frequency characteristics of russian language texts]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Science bulletin of the Novosibirsk state technical university*, 2017, no. 2, pp. 146–162. DOI: 10.17212/1814-1196-2017-2-146-162.

13. Kotov Yu.A., Sanina O.V. Criteria and algorithm for the Russian language text recognition based on the frequency characteristics set. *XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE 2018)*, Novosibirsk, 2018, pp. 175–179. DOI: 10.1109/APEIE.2018.8545877.

14. Kotov Yu.A., Sanina O.V. Znacheniya nekotorykh bigrammnykh kharakteristik russkoyazychnykh tekstov [Importance of some bigram characteristics for Russian language texts]. *Vestnik SibGUTI = The Herald of SibSUTIS*, 2017. no. 4, pp. 24–34.

15. Zhdanov O.N., Kudenkova I.A. *Kriptoanaliz klassicheskikh shifrov* [Cryptoanalisys of classical ciphers]. Krasnoyarsk, Reshetnev Siberian State Aerospace University Publ., 2008. 107 p.

16. Kotov Yu.A., Sanina O.V. Identifikatsiya probela pri neizvestnoi znakovoi kodirovke v russkoyazychnykh tekstakh [Space character identification in Russian language texts with unknown encoding]. *Vestnik SibGUTI = The Herald of SibSUTIS*, 2018, no. 4, pp. 48–60.

17. Smith R.D. Distinct word length frequencies: distributions and symbol entropies. *Glottometrics: Studies in Quantítative Linguistics*, 2012, no. 23, pp. 7–22.

18. Moreno-Sanchez I., Font-Clos F., Corral A. Large-scale analysis of Zipf's law in English texts. *PLoS One*, 2016, vol. 11, no. 1. DOI: 10.1371/journal.pone.0147073.

19. Ferrer-i-Cancho R., Elvevag B. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One*, 2010, vol. 5, no. 3. DOI: 10.1371/journal.pone.0009411.

Для цитирования:

Котов Ю.А., Санина О.В. Формирование псевдослучайных текстов на основе частотных характеристик текстов естественного языка // Сборник научных трудов НГТУ. – 2020 – № 1–2 (97). – С. 113–126. – DOI: 10.17212/2307-6879-2020-1-2-113-126.

For citation:

Kotov Yu.A., Sanina O.V. Formirovanie psevdosluchainykh tekstov na osnove chastotnykh kharakteristik tekstov estestvennogo yazyka [Generating pseudo-random texts based on the frequency characteristics of texts in natural languages]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta = Transaction of scientific papers of the Novosibirsk state technical university*, 2020, no. 1–2 (97), pp. 113–126. DOI: 10.17212/2307-6879-2020-1-2-113-126.