Рассмотрим класс эквивалентности $C_i = \{s^1, \ldots, s^m\}$. Заметим, что для того, чтобы проверить выполнение условия безопасности 1, достаточно проверить пары $(s^1, s^2), \ldots, (s^1, s^m)$. Если $\delta(s^1, x_L) \sim \delta(s^j, x_L)$ и $\delta(s^1, x_L) \sim \delta(s^t, x_L)$, то $\delta(s^j, x_L) \sim \delta(s^t, x_L)$, то есть для пары (s^j, s^t) условие безопасности 1 также выполнено. Таким образом, сложность проверки условия безопасности 1 равна $O(|S| \cdot |\Sigma_L|)$.

В результате, сложность проверки безопасности составляет $O(|S| \cdot |\Sigma|)$.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Afonin Sergey, Bonushkina Antonina. Validation of Safety-Like Properties for Entity-Based Access Control Policies // Advances in Soft and Hard Computing. 2019. С. 259–271.

2. Harrison M.A., Ruzzo W.L., Ullman J.D.. Protection in Operating Systems // Communications of the ACM. 19 №8. 1976. С. 461–471.

3. Moskowitz I.S., Costich O.L.. A classical automata approach to noninterference type problems// Proc. Computer Security Foundations Workshop V. IEEE Press. 1992. С. 2–8.

4. Грушо А.А., Тимонина Е.Е.. Теоретические основы защиты информации // Яхтсмен. Москва. 1996.

УДК 519.722

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

А. Г. Малашина (Россия, г. Москва)
НИУ ВШЭ
e-mail: amalashina@hse.ru
А. Б. Лось (Россия, г. Москва)
НИУ ВШЭ
e-mail: alos@hse.ru

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

A. G. Malashina (Russia, Moscow)
NRU HSE
e-mail: amalashina@hse.ru
A. B. Los (Russia, Moscow)
NRU HSE
e-mail: alos@hse.ru

1. Введение

Исследование вероятностного и статистического распределения слов и n-грамм естественного языка является предметом анализа во многих областях: лингвистике, теории игр, молекулярной биологии. Важную роль корпусный анализ языка играет в вопросах криптографической защиты информации, в том числе в вопросах эффективности ряда криптографических алгоритмов. При исследовании процедур восстановления отдельных участков сообщения по имеющейся информации о вариантах его знаков, основу анализа составляют создаваемые словари различных длин. В связи с этим, особое значение имеет изучение их статистических свойств, проверка полноты и адекватности используемого корпуса [1]. При исследовании языкового корпуса и составлении словарей, одним из основных вопросов является вопрос покрытия словарем всех возможных отрезков текста [2]. При этом проблема покрытия существенно усложняется для флективных языков, таких как французский и немецкий, и в особенности русский, по сравнению с аналитическими языками, такими как английский. Такие языки требуют большего объёма словаря для достижения необходимого покрытия [3]. В работе исследуются две языковые модели: лексическая и n-граммная. В лексической модели языка единицей анализа являются токены, то есть единицы текста, элементы раздельного написания. В n-граммной модели, являющейся частным случаем лексической, рассматриваются последовательности из n символов или слов.

2. Языковой корпус

Корпус - собрание текстов в текстовой форме, используемое для исследования языка с использованием компьютерных технологий. В данной работе был создан специализированный корпус русского языка, который отражает узкую область его употребления. В качестве исходного материала для составления корпуса были использованы новостные статьи последних лет политической тематики. Эти тексты отражают срез состояния современного русского языка, включая разговорный, то есть составляемый корпус является синхроническим.

После создания текстового корпуса осуществляется его нормализация, состоящая из следующих этапов: 1) удаление html-тегов и переформатирование в *.txt; 2) перекодировка; 3) удаление всех сокращений, кроме аббревиатур; 4) удаление имён нарицательных ; 5) фильтрация текста (удаление всех символов, кроме «а-я», «.», «,», « », приведение к нижнему регистру); 6) удаление двойных пробелов, повторяющихся точек и запятых, пробелов перед точками и запятыми.

Метаданные в корпусе отсутствуют, так как их наличие не принципиально для дальнейших целей использования данного корпуса. Созданный корпус должен удовлетворять двум основным критериям: полноте и репрезентативности. Полнота корпуса обуславливается покрытием этого корпуса. Оптимизация покрытия зависит от задач, для которых создаётся этот корпус и словари. Во-первых, покрытие зависит от объёма текстового корпуса, который используется для построения словарей, но с определенного момента эта зависимость становится гораздо менее выраженной, поэтому возможна экстраполяция логарифмическими функциями. Например, для английского языка рост объёма словаря существенно замедляется при размере корпуса от 30 до 50 млн. слов. Во-вторых, оптимальный размер корпуса зависит от источников и новизны данных [3]. В целом, корпус считают насыщенным, когда с увеличением объёма корпуса прекращается резкий рост новых слов [2]. Репрезентативность — это способность корпуса адекватно отражать специфику выбранной предметной области. На основе собранного новостного текстового корпуса в соответствии с рассматриваемой языковой моделью создаются словари, которые впоследствии подвергаются статистическому исследованию.

3. Лексическая модель

Лексические модели языка представляют особый интерес для систем распознавания речи. Актуальность вопроса максимизации покрытия связана с тем, что каждое неизвестное слово (называемое также out-of-vocabulary или OOV) создаёт очередную ошибку в распознавании текущего слова. Более того, каждая такая ошибка способна породить ошибку распознавания следующего слова, создавая «волновой эффект» OOV-слов. В рамках лексической модели языка генерируются словари, единицами которых являются токены, то есть единицы текста как элементы раздельного написания. В таблице ниже представлены значения размеров словарей, создаваемых на основе корпусов различного объёма.

Корпус, симв.	Объем словаря	Эмпирическое покрытие	Теоретическое покрытие
10^4	836	0.45	20.1
10^5	5449	2.74	28.34
10^6	31895	13.78	39.68
10^7 с нарицат.	163091	70.63	48.29
10^7 без нарицат.	125162	64.45	44.98
10^8	*180000*	-	*59.28*
10^9	*230000*	-	*68.84*

Поскольку скорость роста объёма словарей в зависимости от размера корпуса близка к скорости роста логарифмической функции, проводится следующая экстраполяция:

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

Прогнозируемые значения объёма словарей обозначены в таблице курсивом. Графически данная зависимость отражена на рисунке 1.

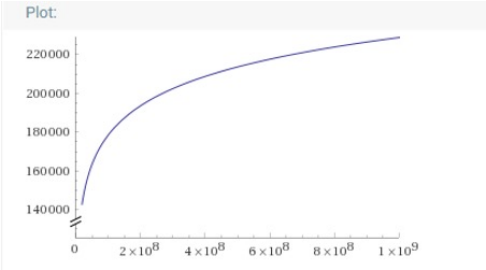

Рис. 1: Размер словаря в зависимости от объёма корпуса.

4. Закон Ципфа.

Для проверки качества и естественности созданного русскоязычного корпуса (объёмом 10^7 символов) проводится проверка соответствия закону Ципфа. В соответствии с данным законом если все слова в корпусе упорядочить по убыванию частоты их встречаемости, то частота использования слов окажется обратно пропорциональной их порядковому рангу [4]. Результаты представлены на рисунке 2.

Как наглядно демонстрирует график, эмпирически полученные значения лишь незначительно отклоняются от идеальных значений закона Ципфа. Поэтому можно заключить, что собранный корпус, в целом удовлетворяет данному закону и является достаточно естественным и полным в рамках лексической модели.

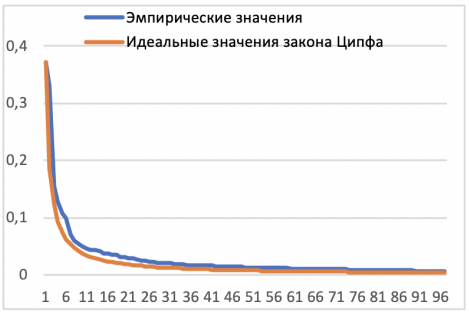

Рис. 2: Закон Ципфа.

4. Лемматизация

Лемматизация - процесс обработки языкового корпуса, в результате которого все входящие в него слова приводятся к словарной форме. Для исследуемых корпусов проводится лемматизация, после чего оценивается величина покрытия. Сравнительные результаты приведены в таблице.

Корпус	Покрытие (оригинал)	Покрытие (лемматизация)
10^6	13.78	16.16
10^7 с нарицат.	70.63	76.45
10^7 без нарицат.	64.45	66.77

Результаты наглядно демонстрируют, что величины покрытия лемматизированных корпусов выше покрытия оригиналов. Это связано с тем, что лемматизация значительно снижает флективность языка, приближая его к аналитическому [3]. Например, в исследуемом корпусе присутствует слово «репрезентативные», а в тестовом корпусе только слово «репрезентативный». То есть одно и то же слово присутствует в корпусах в разной форме. Но с точки зрения автоматической обработки, точно сопоставляющей токены, это разные слова. Это снижает величину покрытия. Лемматизация устраняет данный недостаток и снижает объём словаря.

5. N-граммная модель

В n-граммной модели единицами корпуса (словаря) считаются последовательности из n символов. Лексические модели являются подмножеством n-граммных моделей языка. Вопросы n-граммного покрытия представляют интерес в системах распознавания речи для максимизации производительности системы. Но, особое значение, n-граммные модели имеют для вопросов криптографии, так как в шифрованном тексте границы между словами неизвестны, и подбор в этом случае по лексическому словарю невозможен или крайне затруднителен. Если какая-либо n-грамма отсутствует в словаре, то языковая модель может опираться на n-граммы более низкого порядка, но они могут оказаться неуместны для текущей задачи. Именно поэтому ошибки распознавания гораздо чаще происходят в рамках n - граммной модели языка. На основе собранных корпусов проводится оценка покрытия и подсчитывается энтропия n-грамм длиною i символов:

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

где i - длина n-граммы, а N_i - количество n-грамм в слове длиною i символов. Рассматриваются n-граммы длиною 10, 15, 20 и 25 символов. Результаты исследования приведены в таблице.

Длина n-граммы	Объем словаря		Энтропия	
	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

Значения энтропии n-грамм близки к реальным значениям для русского языка. Величины покрытия представлены в таблице ниже.

Длина n-граммы	Эмпирическое покрытие		Теоретическое покрытие	
	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

Как видно, покрытие словаря n-грамм значительно ниже лексического покрытия. Это подтверждает свидетельства о том, что исследование n-граммной модели гораздо более сложное, а оптимизация покрытия словаря требует сверхбольшого языкового корпуса.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Малашина А. Г. Алгоритм восстановления отдельных частей сообщения по информации о возможных значениях его знаков, Материалы конференции // Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского. – Москва, 2019. С. 215-217.

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