seti-5ab251a455876bb026a5bb79 (Дата обращения: 15.12.2019).

3. Branco P., Mateus P. A Traceable Ring Signature Scheme Based on Coding Theory // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019. Vol. 11505 LNCS. P. 387–403 [Электронный ресурс]. URL: https://eprint.iacr.org/2019/324.pdf (Дата обращения: 15.12.2019).

4. Кольцевые подписи и их приложения - CryptoWiki [Электронный ресурс]. URL: http://cryptowiki.net/index.php?title=Кольцевые_подписи_и_их_приложения (Дата обращения: 15.12.2019)

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЯЗЫКОВЫХ МОДЕЛЕЙ РУССКОГО ЯЗЫКА НА ОСНОВЕ НОВОСТНОГО ТЕКСТОВОГО КОРПУСА

А.Г. Малашина
Национальный исследовательский университет "Высшая школа экономики", аспирантская школа по техническим наукам

Аннотация

В работе проводится статистический анализ свойств лексических и n-граммных моделей русского языка на основе новостного текстового корпуса.

Создан специализированный корпус из новостных статей последних лет политической направленности, отражающий узкую область употребления языка. Составлены словари токенов и n-грамм, найдены величины покрытия этих словарей, а также значения энтропии. Проведена лемматизация исходного текстового корпуса и экстраполяция роста объёма словарей в зависимости от увеличения размера корпуса.

Введение

Исследование вероятностного и статистического распределения слов и n-грамм естественного языка является предметом анализа во многих областях: лингвистике, теории игр, молекулярной биологии и криптографии. Важную роль корпусный анализ языка играет в вопросах защиты информации, в том числе в вопросах бесключевого восстановления открытого текста. При исследовании возможности восстановления отдельных участков сообщений, зашифрованных неполной или неравновероятной гаммой, основу анализа составляют создаваемые словари, поэтому бесспорную важность имеет изучение их статистических свойств и проверка полноты и адекватности используемого корпуса [1].

При исследовании языкового корпуса и составлении словарей одной из основных проблем является проблема покрытия, то есть выбор словаря так, чтобы в исследуемых текстах присутствовало как можно меньше неизвестных слов [2]. При этом проблема покрытия существенно усугубляется 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 покрытия, аналогичному английскому [3]. Необходимо создать корпус, покрытие которого будет максимально приемлемо для рассматриваемых задач.

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.

Языковой корпус

Корпус - собрание текстов в текстовой форме, используемое для исследования языка с использованием компьютерных технологий [4]. 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 [4, 5]. 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 (данный этап проводится не для всех корпусов);
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 [6].

Репрезентативность – это способность корпуса адекватно отражать специфику выбранной предметной области [4].

На основе собранного новостного текстового корпуса в соответствии с рассматриваемой языковой моделью создаются словари, которые впоследствии подвергаются статистическому исследованию.

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.

Таблица 1. Объём и покрытие

Корпус, симв.	Объем словаря	Эмпирическое покрытие, %	Теоретическое покрытие, %
$\mathbf{10^4}$	836	0,45	20,1
$\mathbf{10^5}$	5449	2,74	28,34
$\mathbf{10^6}$	31895	13,78	39,68
$\mathbf{10^7}$ с нарицат.	163091	70,63	48,29
$\mathbf{10^7}$ без нарицат.	125162	64,45	44,98
$\mathbf{10^8}$	*~180000*	-	-
$\mathbf{10^9}$	*~230000*	-	-

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

$$21910{,}5\ log(0{,}000034344x)$$

Прогнозируемые значения объёма словарей обозначены в Таблиц курсивом.

Графически данная зависимость отражена на рисунке ниже:

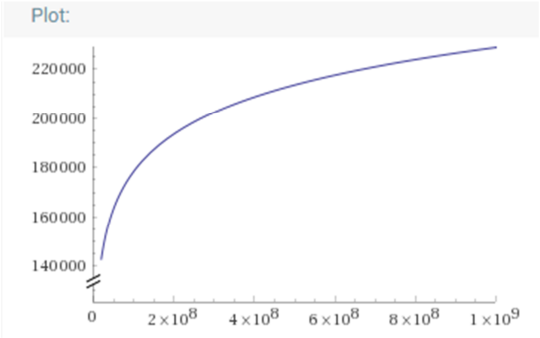

Рис.1. Размер словаря в зависимости от объёма корпуса

Закон Ципфа

Для проверки качества и естественности созданного русскоязычного корпуса (объёмом 10 млн. 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 [7].

Результаты представлены на рисунке ниже:

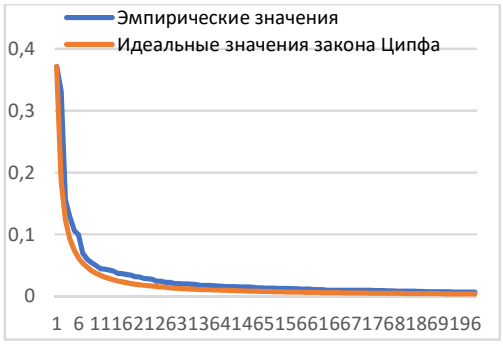

Рис.2. Закон Ципфа

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.

Покрытие корпуса

Для оценки полноты рассматриваемого корпуса требуется оценить величину покрытия его словаря. Существуют различные подходы к оценке покрытия. В данном исследовании были рассмотрены эмпирическая и теоретическая оценки.

Эмпирическая оценка проводится с помощью тестового корпуса, не являющегося частью проверяемого. Величина покрытия — это доля токенов тестового корпуса, покрытая словарём исходного корпуса. Теоретическая оценка основывается на количестве уникальных токенов словаря, то есть единиц словаря, встречающихся только один раз:

$$\widetilde{N} = \frac{N}{1-\frac{n}{N}},$$

где N – исходный объём словаря, n – число токенов, встречающихся один раз.

Предполагается, что полученная таким образом величина – это приблизительный размер полного словаря. На основании теоретически найденного объёма словаря оценивается величина покрытия как отношение практической величины объема к теоретической. Покрытие выражается в процентах.

Значения эмпирического и теоретического покрытия указаны в таблице 1. С увеличением объёма корпуса величина покрытия возрастает. Для маленьких корпусов теоретически найденное покрытие значительно превышает экспериментальное. Такое различие вскрывает недостатки подхода, основанного на количестве однократно встречаемых единиц словаря, особенно для корпусов небольшого объёма. Для больших корпусов эмпирическое покрытие выше теоретического, что отчасти может быть объяснено тем, что среди однократно встречающихся токенов в словаре могут присутствовать слова, написанные с опечатками, редкие имена и фамилии, специфические названия, узкоспециализиро-ванная лексика и редко употребляемые слова. В качестве примера, обличающего подобный недостаток теоретической оценки покрытия, можно привести Брауновский корпус английского языка. Более половины слов в данном корпусе встречаются лишь однажды. Следовательно, относительно теоретической оценке его покрытие составляет менее 50% несмотря на то, что объём корпуса около 1 млн. слов [4].

Лемматизация

Лемматизация – процесс обработки языкового корпуса, в результате которого все входящие в него слова приводятся к словарной форме [4].

Для исследуемых корпусов проводится лемматизация, после чего оценивается величина покрытия. Сравнительные результаты приведены в таблице 2.

Таблица 2. Покрытие

Корпус	Покрытие (оригинал), %	Покрытие (леммат-из.), %
$\mathbf{10^6}$	13,78	16,16
$\mathbf{10^7}$ с нарицат.	70,63	76,45
$\mathbf{10^7}$ 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.

Процесс лемматизации может оказаться крайне полезен, например, для генерации словарей на основе корпусов. Но для некоторых других задач, таких как распознавание речи или бесключевое восстановление открытого текста, словаря (в более общем смысле), содержащего только словарные формы, может оказаться недостаточно. В таком случае расширение словаря может производиться в два этапа: с помощью лемматизации, а затем процесса, обратного ей. Например, в корпусе встречается только слово «языковые». Леммой данного слова является слово «языковой». В словарь могут добавляться все орфографические формы данного слова: «языковая», «языковое», «языковых», «языковым» и т.д. Более того, словарь может быть расширен ещё больше, если добавлять в него также однокоренные токены других частей речи, например, «язык». Но такой подход требует наличия промежуточного автоматизированного блока, умеющего производить орфографические изменения слов.

Ключевые слова

На основе корпуса объёмом 1 млн. символов определяются ключевые слова, то есть слова, встречающиеся чаще всего. Результаты представлены в двух вариантах: для исходного и лемматизированного корпуса. Топ-10 слов приведен в таблице 3.

Таблица 3. Ключевые слова

№	Лемматизированный корпус (1 млн. символов)	Количество	Оригинальный корпус (1 млн. символов)	Количество
1.	Россия	663	России	420
2.	страна	656	Сирии	290
3.	политический	573	США	270
4.	Сирия	351	время	228
5.	год	348	страны	208
6.	человек	342	Россия	173
7.	российский	339	политической	160
8.	один	332	власти	146
9.	много	328	против	142
10.	время	320	РФ	133

Ключевые слова подчеркивают специфичность и специализированность корпуса, собранного на основе современных новостных статей на тему политики.

N-граммная модель

В n-граммной модели единицами корпуса (словаря) считаются последовательности из n символов. Лексические модели являются подмножеством n-граммных моделей языка.

N-граммное покрытие также представляет интерес в системах распознавания речи для максимизации производительности системы. Но особое значения n-граммные модели имеют для вопросов информационной безопасности, 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 [8].

Анализ покрытия словаря n-грамм усложняется из-за значительно меньшей частоты n-грамм по сравнению с наименее частыми словами в словаре. По оценкам сборника североамериканских новостных деловых статей (NAB), который на данный момент является самым тщательно исследованным языковым корпусом, чтобы оптимизировать охват биграмм (два слова), требуется корпус объёмом от 100 до 200 млн. слов. Собрать текстовой корпус такого объема представляется довольно трудной задачей. А с увеличением n проблема оптимизации покрытия только ухудшается.

Ещё больше осложняет ситуацию существующее потенциальное взаимодействие между покрытием n-грамм и эволюцией языка. Накопление слов из соответствующего источника, очевидно, занимает время, в течение которого языковые шаблоны могут меняться, ухудшая адекватность более старых данных. Рассматривая язык как нестационарный стохастический источник, Розенфельд постулировал следующий принцип: никогда нельзя определить одновременно и степень, и временные рамки языкового явления. Как следствие, он пришел к выводу, что невозможно обнаруживать преходящие и редкие лингвистические события [2].

На основе собранных корпусов проводится оценка покрытия и подсчитывается энтропия n-грамм длиною i символов:

$$H_i = \frac{log_2 N_i}{i},$$

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 их исследования приведены в таблице 4.

Таблица 4. Объём и энтропия

Длина n-граммы	Объем словаря		Энтропия (бит/симв)	
	10^6	10^7	10^6	10^7
10	692698	6217191	1,94	2,26
15	889939	9482897	1,32	1,55
20	941634	10372296	0,99	1,17
25	958168	10629589	0,79	0,93

Значения энтропии n-грамм близки к реальным значениям для русского языка.

Величины покрытия представлены в таблице 5.

Таблица 5. Покрытие

Длина n-граммы	Эмпирическое покрытие, %		Теоретическое покрытие, %	
	10^6	10^7	10^6	10^7
10	4,32	39,71	16,07	19,95
15	1,1	12,03	7	7,21
20	0,28	3,12	4,17	3,27
25	0,07	0,84	3,18	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.

Заключение

В данной работе собран и обработан текстовой корпус, основанный на новостных статьях последних лет. На базе созданного русскоязычного корпуса сгенерированы слова-

ри для разных языковых моделей. Рассмотрены различные подходы определения величины покрытия созданных словарей, позволяющей оценить полноту собранного корпуса. С помощью выделения ключевых слов проверена репрезентативность исследуемого корпуса как его способность адекватно отражать специфику выбранной политической тематики. С помощью закона Ципфа была оценена естественность и качество рассматриваемого новостного корпуса. Были подсчитаны значения энтропии n-грамм различной длины. Найденные значения энтропии приближены к реальным значениям энтропии русского языка, что говорит о том, что корпус объёмом 10 млн. символов удовлетворяет основным выдвигаемым критериям - критерию полноты и адекватности - поэтому собранный корпус может быть использован как основа для дальнейших исследований.

Список литературы:

1. А. Г. Малашина, «Алгоритм восстановления отдельных частей сообщения по информации о возможных значениях его знаков,» в Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского. Материалы конференции, Москва, 2019.
2. R. Rosenfeld, «Optimizing lexical and n-gram coverage via judicious use of linguistic data» в Proceedings of the Fourth European Conference on Speech Communication and Technology, Madrid, 1995.
3. J. R. Bellegarda, «Robustness in Statistical Language Modeling» в Robustness in Language and Speech Technology , Springer Science+Business Media Dordrecht, 2001, pp. 104-106.
4. Т. М. Волосатова, Информатика и лингвистика: учеб. пособие, Т. М. Волосатова и Н. В. Чичварин, Ред., ИНФРА-М, 2018, p. 196 с.
5. И. С. Кипяткова, «Исследование статистических n-граммных моделей языка для распознавания слитной русской речи со сверхбольшим словарем» в Анализ разговорной русской речи , Санкт-Петербург, 2010.
6. А. Б. Викторов, С. Г. Грамницкий, С. С. Гордеев, М. В. Ескевич и Е. М. Климина, «Универсальная методика подготовки компонентов обучения систем распознавания речи», Речевые технологии, pp. 39-56, Февраль 2009.
7. A. Gelbukh и G. Sidorov, «Zipf and Heaps Laws' Coefficients Depend on Language» в Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2001.
8. L. Chase, R. Rosenfeld и W. Ward, «Error-responsive modifications to speech recognizers: negative n-grams» в Third International Conference on Spoken Language Processing, Yokohama, 1994.

ПОДХОД К КЛАССИФИКАЦИИ ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

А.А. Спирин
Академия ФСО России

Аннотация

Для решения задачи обнаружения и блокировки DLP-системами канала передачи зашифрованных или сжатых данных предлагается использовать статистические свойства данных в двоичном представлении. Для построения классификатора используется алгоритм формирования дерева решений.

Введение

В настоящее время, согласно отчетам информационно-аналитических агентств, возросло количество утечек конфиденциальных данных [1]. Более чем в 55% случаев утечка конфиденциальных данных произошла по вине внутреннего нарушителя.

Для предотвращения утечек применяются различные программные и аппаратные средства, одними из них являются системы предотвращения утечек конфиденциальных данных – data leakage prevention systems (DLP-системы).

В настоящее время DLP-системы используют методы, выполняющие сканирование передаваемой информации с целью обнаружения определенных слов, фраз, регулярных выражений. Также осуществляется анализ контекста передаваемых данных: ip-адрес, порт отправителя/получателя, длина пакетов, наличие определенных флагов и др. [2-9] Однако существуют способы обхода указанных методов защиты, например, шифрование или сжатие [10,11].

Таким образом, одной из возможных причин утечек конфиденциальных данных может являться канал передачи информации в зашифрованном или сжатом виде, организуемый внутренним нарушителем. Канал передачи информации не блокируется DLP-системами по причине отсутствия в нем сигнатур, по которым происходит обнаружение конфиденциальных данных.

С целью блокировки передачи зашифрованных данных необходим подход, опирающийся на статистические свойства информации в двоичном представлении. Предлагается использовать методы машинного обучения для построения классификатора, обнаруживающего зашифрованные и сжатые данные.

Для решения задачи построения классификатора, необходимо выбрать признаковое пространство, являющееся множеством двоичных подпоследовательностей, которое позволит выполнять классификацию псевдослучайных последовательностей с точностью более 0.95.

Исходные данные для эксперимента

Для построения классификатора была сформирована выборка из 9000 файлов в двоичном представлении, относящихся к 4-м классам:

1. Зашифрованные криптоалгоритмами AES, 3DES, RC4, Camellia [12].
2. Архивы RAR, ZIP [13].
3. Зашифрованные архивы RAR, ZIP [13].
4. Сформированные генератором псевдослучайных чисел (утилитой urandom операционной системы семейства Unix [14]).

Далее под псевдослучайными последовательностями (ПСП) понимается двоичное представление файлов из выборки.

Построение признакового пространства на основе подсчета частоты встречаемости двоичных подпоследовательностей длины N бит

Было сделано предположение о том, что в качестве признакового пространства могут быть использованы частоты появления независимых битовых подпоследовательностей различной длины N без полного перекрытия подпоследовательностей. Например, для последовательности S=00011011 частота вхождения подпоследовательностей длины N=2 бит представлена в таблице 1.

Таблица 1. Пример расчета частоты подпоследовательностей длины N=2 бит

Подпоследовательность	Количество	Частота
00	1	0,14
01	2	0,29
10	1	0,14
11	2	0,29

Таким образом признаковое пространство представляет собой значение частот появления подпоследовательностей длины N бит в исследуемых данных. Частота появления