Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Малашина Анастасия Геннадьевна

Разработка методов и алгоритмов анализа характеристик текстов на естественном языке на основе теоретико-информационного подхода и применение их к задачам защиты информации

РЕЗЮМЕ ДИССЕРТАЦИИ

на соискание ученой степени кандидата
технических наук

Научный руководитель:
кандидат технических наук, доцент
Лось Алексей Борисович

Москва – 2024

Оглавление

Введение

Постановка проблемы и актуальность исследования

Одним из аспектов обеспечения информационной безопасности является применение теоретико-информационного подхода к анализу алгоритмических методов защиты информации. В частности, в рамках данного подхода рассчитываются информационные характеристики и параметры источника сообщений, от которых, в том числе, зависят вопросы его защиты. Данная задача весьма актуальна, однако достаточно сложна и, несмотря на имеющуюся теоретическую базу, в указанной предметной области проводилось мало прикладных исследований.

Учитывая актуальность данного направления, в диссертационной работе разрабатываются методы и алгоритмы анализа характеристик различных текстов на естественных языках, таких как энтропия коротких s-грамм, мощностные характеристики и покрытие словарей s-грамм с использованием теоретико-информационного подхода.

Результаты анализа информационных характеристик текстов применимы для решения задач информационной безопасности и развития методов анализа и синтеза современных алгоритмов защиты информации. В частности, они могут быть использованы для построения методов восстановления участков сообщений на основе теоретико-информационного подхода, а также для исследования эффективности алгоритмов аутентификации на основе парольной защиты.

Например, в ряде случаев возникают ситуации, когда относительно знаков неизвестного текстового сообщения на выходе канала связи появляется информация о возможных значениях исходных символов.

Обладая некоторой информацией о знаках неизвестного текста и заранее ограничивая множество входных сообщений (например, с помощью закрытых словарей), можно реализовать восстановление исходного сообщения или его отдельных s-грамм [12, 15].

Объектом исследования в настоящей работе являются тексты на естественном языке. **Предметом исследования** являются методы и алгоритмы анализа информационных характеристик данных текстов.

Актуальность данного исследования обусловлена необходимостью изучения возможности применения теоретико-информационного и алгоритмического подходов к ряду задач информационной безопасности и развитием методов анализа и синтеза современных алгоритмов защиты информации.

Актуальность и значимость исследования подтверждается п. 1.4 «Разработка теоретико-информационных методов анализа систем передачи, хранения и защиты информации» паспорта области науки «Инженерные науки и прикладная математика» [11], утвержденного Ученом советом НИУ ВШЭ от 2 февраля 2018 г., а также недавними исследованиями, проводимыми в данном направлении [1].

Цели и задачи диссертационной работы

Основной целью данной диссертационной работы является разработка методов и алгоритмов анализа характеристик текстов на естественном языке на основе теоретико-информационного подхода.

Для решения соответствующей цели были поставлены следующие **задачи**:

1. Разработать алгоритмы нормализации текстовых корпусов и построения словарей s-грамм. Сформировать различные текстовые корпуса и построить словари s-грамм.
2. Разработать математическую модель покрытия словарей s-грамм. Получить численные оценки покрытия словарей.
3. Разработать метод оценки энтропии s-грамм. Получить численные оценки энтропии коротких s-грамм.
4. Исследовать границы допустимой многозначности восстановления s-грамм сообщения.

5. Разработать метод восстановления участков входных сообщений с использованием теоретико-информационного подхода. Получить численные оценки доли восстановленных s-грамм сообщения.

Степень разработанности темы

Теоретико-информационные аспекты защиты информации подробно описаны в монографиях многих мировых ученых, в том числе К. Шеннона [7].

Отдельные случаи построения методов восстановления текстовой информации на выходе канала связи с использованием теоретико-информационного подхода в настоящее время исследуются А. В. Бабашом [1], а также исследовались В. М. Деундяком, Е. А. Пашковой [10]. В этих работах применяется метод «чтения в колонках» для полного восстановления выходного сообщения при наличии информации на выходе канала связи, которая значительно ограничивает число неизвестных символов сообщения (с вероятностью их появления близкой к 1).

Согласно ранее проведенным исследованиям [15], полное восстановление сообщения возможно, если количество предполагаемых значений исходных символов ограничено (например, для русского языка – не более 16 при алфавите в 32 символа), при этом вероятность появления среди них истинного знака близка к единице. Тогда для каждого входного символа можно зафиксировать множество его возможных значений (например, наиболее вероятных). Восстановление исходного текста представляет собой поиск варианта из множества входных сообщений среди всех возможных комбинаций. Такой подход может привести к потере истинного варианта восстановления, вероятность которой оценивается исходя из введенной теоретико-вероятностной модели исследуемого процесса. С увеличением количества значений на выходе канала связи восстановление исходного текста становится затруднительным из-за значительного количества возможных вариантов восстановления и неопределенности выбора, возникающей в процессе перебора. Таким образом, когда количество символов на выходе канала невелико, вариант восстановления из множества входных сообщений

может быть построен с допустимой многозначностью так как все остальные комбинации окажутся текстом случайной структуры. Однако по мере увеличения числа значений такой подход приводит к нахождению множества вариантов, превышающих допустимую многозначность. В этом случае невозможно определить, какой из найденных текстов является исходным сообщением, не обладая дополнительной информацией. То есть полное восстановление сообщения становится невозможным. Однако вопрос возможности восстановления отдельных s-грамм сообщения остается открытым.

Основой теоретико-информационного подхода к информационной безопасности является понятие энтропии. Текст и язык также являются системами, обладающими информационной энтропией. Более того, энтропия текстов на естественном языке значительно ниже максимальной энтропии алфавита [17]. Поскольку количество допустимых s-грамм в языке значительно меньше, чем запретных, такой подход значительно снижает сложность восстановления по сравнению с полным перебором. Аналогичный подход, например, может быть использован при подборе паролей [4].

Существуют различные методы оценки энтропии s-грамм текстов. Наиболее известным среди них является метод Шеннона [8], основанный на угадывании возможного продолжения текста по известному отрывку. Метод позволяет получить нижнюю и верхнюю оценки. Вместо метода угадывания, можно использовать представление текста с помощью цепи Маркова глубины *s*, чтобы приблизительно оценить вероятности s-грамм и использовать формулу Шеннона для расчета информационной энтропии [8].

В данной диссертационной работе используется метод определения энтропии s-грамм, основанный на составлении словарей, идеи которого восходят к комбинаторного подходу Колмогорова [5] и предлагаются методы асимптотической оценки покрытия создаваемых словарей s-грамм.

Научная новизна исследования

1. Разработаны алгоритмы нормализации текстовых корпусов и построения словарей s-грамм, позволяющие сформировать прикладную базу для проведения анализа характеристик текстов на естественном языке на основе теоретико-информационного подхода.
2. Предложены и исследованы математические модели покрытия словарей s-грамм, основанные на теоретико-информационных свойствах текстов на естественном языке.
3. Разработан метод оценки энтропии коротких s-грамм, основанный на мощностных характеристиках словарей s-грамм.
4. Проведены исследования границ допустимой многозначности восстановления s-грамм сообщения.
5. Разработан метод восстановления участков входных сообщений с использованием теоретико-информационного подхода.

Теоретическая и практическая значимость исследования

Теоретическая значимость данного диссертационного исследования заключается в развитии теоретико-информационного и алгоритмического подходов для решения задач информационной безопасности и развитии методов анализа и синтеза современных алгоритмов защиты информации.

Практическая значимость результатов диссертационной работы заключается в следующем:

1. Разработанные методы и алгоритмы позволяют получать численные оценки информационных характеристик естественных языков.

2. Разработанный метод восстановления участков входных сообщений позволяет оценить долю информации, которую возможно восстановить с использованием теоретико-информационного подхода при заданных мощностных характеристиках.

3. Методы построения словарей и оценки энтропии коротких s-грамм могут быть использованы для анализа стойкости алгоритмов аутентификации, построенных на основе парольных систем.

Методы исследования

В диссертации использованы методы статистического анализа, предельные теоремы теории вероятностей, включая центральную предельную теорему и теоремы о распределении (в том числе, условных и многомерных) нормальных, биномиальных и гипергеометрических величин, неравенство Берри-Эссена. Использованы основополагающие понятия из теории информации (дискретный канал связи, энтропия дискретного стационарного источника, вторая теорема Шеннона и др.). Для создания и проведения исследования текстовых корпусов и словарей s-грамм разработана программная реализация на языке C++ и применены теоретические аспекты корпусной лингвистики, в том числе принципы о полноте и репрезентативности составляемых корпусов.

Для исследования возможности восстановления отдельных s-грамм сообщения разработана программная реализация алгоритма восстановления с помощью методов объектно-ориентированного программирования на языке C++. Для получения ряда численных оценок и экстраполяции эмпирических результатов использованы средства компьютерной алгебры (численные методы) в составе программного пакета Wolfram Mathematica.

Личный вклад

Все результаты и положения, выносимые на защиту, получены автором лично. В ходе диссертационного исследования разработаны алгоритмы нормализации текстовых корпусов и построения словарей s-грамм, позволяющие сформировать прикладную базу для проведения анализа характеристик текстов на естественном языке на основе теоретико-информационного подхода. Разработаны подходы, которые позволяют оценивать покрытие словарей s-грамм и энтропию коротких s-грамм. Проведены экспериментальные исследования ряда информационных характеристик русского и английского языков: объемов словарей s-грамм для корпусов различных стилей и размеров, покрытия словарей, энтропии s-грамм.

Разработан метод восстановления участков входных сообщений с использованием теоретико-информационного подхода. Исследована возможность восстановления отдельных s-грамм сообщения по заранее сформированному словарю в случае, когда полное восстановление невозможно. Исследована проблема допустимой многозначности восстановления s-грамм сообщения, получены теоретические и экспериментальные оценки границ допустимой многозначности восстановления указанных s-грамм. Проведены экспериментальные исследования по восстановлению подходящих s-грамм сообщения в рамках определенной теоретико-вероятностной модели появления значений символов на выходе канала. Проведены исследования по теоретической оценке вероятностных распределений, возникающих в задаче восстановления отдельных s-грамм сообщения.

Содержание работы

Диссертация состоит из глоссария, введения, 3-х глав, заключения, списка литературы, приложения и содержит 116 страниц, 45 таблиц и 20 рисунков.

В **первой главе** проводится анализ известных подходов в части исследования информационных характеристик языков, в том числе зарубежных стандартов по оценки энтропии случайной последовательности. Приведен обзор языковых моделей языка, которые наиболее широко используются на практике для описания текстов на естественном языке. Рассмотрены s-граммные модели языка, проблемы оценки вероятностей s-грамм в текстовом корпусе и проблема внесловарных (OOV) s-грамм. Проанализированы методы увеличения покрытия словарей s-грамм. Рассмотрены оценки размеров английских корпусов, необходимые для достижения оптимального покрытия. Исследованы марковские модели в качестве приближенных моделей естественного языка, описаны недостатки данных моделей для описания текстов на естественном языке.

Проведен обзор методов оценки энтропии случайной последовательности из американского стандарта NIST SP 800-90B и белорусского стандарта МИ.10127.10.02. Установлено, что анализируемые методы имеют ряд ограничений для использования их в целях оценки энтропии коротких s-грамм текстов на естественном языке, так как язык, как правило, обладает высокой структурной предсказуемостью и содержит много повторений и общепринятых фраз. Кроме того, методы оценки энтропии из стандартов обычно предназначены для анализа больших объемов данных и требуют большой выборки для статистической значимости, что делает их трудноприменимыми для анализа энтропии коротких текстов. Выявлена необходимость в разработке метода оценки энтропии коротких s-грамм.

Также приведен обзор известных подходов в части оценивания предельной величины энтропии для текстов на естественном языке. Рассмотрены подходы Шеннона и Колмогорова, проводимые ими эксперименты. Проанализирован также ряд других исследований по оценке предельной энтропии для английского и русского языков, которые корректируют и уточняют ранее полученные данные. Ниже (см. Таблица 1) приведен анализ известных оценок.

Таблица 1 – Предельные значения энтропии H_0 (бит / символ)

Английский язык			
Автор	*Корпус*	*Алфавит*	*Энтропия, H_0*
Шеннон [8], 1951	Один литературный текст (Dumas Malone «Jefferson the Virginian»)	26 букв и пробел	0,6 – 1,3
Браун [2], 1992	Общеязыковой (Брауновский корпус)	все печатные символы ASCII (95 символов)	1,75
Тихан и Клири [9], 1996	Один литературный текст (Dumas Malone «Jefferson the Virginian»)	26 букв и пробел	1,46
Контояннис [6], 1997	Литературные тексты (4 романа Дж. Остин)	все печатные символы ASCII (95 символов)	1,77
Калин [3], 2020	Поэзия Д. Г. Байрона	26 букв, 10 цифр, пробел и др. знаки (всего 63 символа)	1,37
Проект №338 МИЭМ [18], 2021	Литературные тексты	26 букв, пробел, точка, запятая	1,32-1,72
Русский язык			
Автор	*Корпус*	*Алфавит*	*Энтропия, H_0*
Пиотровский [14], 1968	Язык в целом	32 буквы, пробел	0,87-1,37

	Литературные тексты	32 буквы, пробел	0,87-1,19
	Публицистика и научно-деловая речь	32 буквы, пробел	0,59-0,83
Колмогоров [5], 1968	Один литературный текст (Аксаков С. Т. «Детские годы Багрова-внука»)	31 буква и пробел	1,0-1,2
Проект №338 МИЭМ [18], 2021	Литературные тексты	32 буквы, пробел, точка, запятая	1,49-1,89

Источник: результаты исследований Шеннона (1951), Брауна (1951), Тихана (1996), Р. Г. Пиотровского (1968), Колмогорова (1968) и др.

Во **второй главе** разрабатываются методы и алгоритмы анализа информационных характеристик текстов на русском и английском языках: построение корпусов и словарей s-грамм, разработка методов оценки энтропии как отдельных s-грамм, так и длинных текстов, разработка моделей покрытия словарей s-грамм и методов оценки допустимой многозначности, появляющейся при восстановлении отдельных участков сообщений. Разработан алгоритм нормализации корпусов текстов и его программная реализация на языке C++. Создано 4 корпуса текстов для исследования: газетно-публицистический корпус русского языка, художественный (литературный) корпус русского языка, общеязыковой корпус русского языка и общеязыковой корпус английского языка. В рамках s-граммной модели языка разработан алгоритм формирования словарей s-грамм из созданных корпусов и его программная реализация на языке C++. Сформированы словари 10-грамм, 15-грамм, 20-грамм и 25-грамм для корпусов различных стилей и языков. Проведено сравнительное экспериментальное исследование объемов словарей s-грамм для 4 корпусов. Проанализирована зависимость объема словаря от размера текстового корпуса. С помощью системы Wolfram Mathematica проведена интерполяция данной зависимости для различных корпусов.

Введена модель источника сообщений на основе s-грамм, в рамках которой рассмотрены два теоретических подхода к оценке покрытия созданных словарей, то есть доли допустимых в языке s-грамм, попавших в словарь. Первая теоретическая модель покрытия основана на количестве однократно встречаемых s-грамм. Вторая теоретическая модель покрытия основана на

нахождении распределения числа «пустых ящиков» в классической задаче о размещении. Для второй модели покрытия проведён асимптотический анализ формулы в случае $s \to \infty$. Установлено, что в рамках данной модели число допустимых s-грамм языка, не присутствующих в словаре, распределено асимптотически нормально. Приведены численные оценки покрытия словарей s-грамм для 4 корпусов. Реализован экспериментальный метод оценки покрытия словарей s-грамм для апробации предложенных теоретических моделей покрытия. Проанализировано, как покрытие словарей влияет на величину ошибки I рода при восстановлении текста по словарю.

Разработан метод оценки энтропии коротких s-грамм, основанный на мощностных характеристиках словарей s-грамм и значении покрытия данных словарей. Идеи метода восходят к комбинаторному подходу А. Колмогорова к оценке количества информации, основанному на статистических характеристиках текста, без привлечения вероятностной модели. Получены численные оценки энтропии коротких s-грамм для разных корпусов. Разработан метод экстраполяции значений оценок энтропии s-грамм с ростом s. Проведена экстраполяции оценок энтропии коротких s-грамм. Получены численные оценки предельной энтропии для разных корпусов.

Исследованы границы многозначности восстановления s-граммы. Предложен подход к оценке, основанный на подсчете среднего числа фрагментов слов, находящихся в s-грамме. Выведена параметрическая форма теоретической оценки. Проведено экспериментальное исследование границ многозначности восстановления s-граммы для проверки теоретических предположений.

Выводы по главе 2

1. По итогам экспериментального исследования мощностных характеристик словарей s-грамм на разном текстовом материале можно сделать следующие выводы:
   - Размеры словарей s-грамм для русского и английского языков различаются: английский обладает меньшим словарем. Однако с ростом s данная разница сглаживается (для 10-грамм она составляет около 25%, для 25-грамм – менее 6%). Данный факт объясняется снижением насыщенности корпуса в рамках s-граммной модели языка с ростом s.
   - Для разных стилей текстов в рамках одного языка наблюдаются некоторые различия в размерах словарей. Например, для русского языка исследовались газетно-публицистические тексты (новости и статьи политической тематики), художественные тексты (русская литература XIX-XX вв.) и тексты смешанных стилей. Для всех длин s-грамм наименьшим словарем обладают газетно-публицистические тексты, наибольшим – художественные (для 10-грамм разница в размерах словарей составила более 15%). Однако с ростом s разница сглаживается, что также объясняется снижением насыщенности корпуса в рамках s-граммной модели языка с ростом s.
2. Результаты исследования зависимости между размером исходного корпуса и объемом соответствующих словарей позволяют заключить следующее:
   - Для английского словаря 10-грамм удается наблюдать некоторое снижение скорости роста словаря при линейном увеличении объёма исходного корпуса. Для других длин s-грамм наблюдается практически линейная зависимость между размером словаря и объёмом корпуса.
   - Для русского языка для всех длин s-грамм наблюдается близкая к линейной зависимость между размером словаря и объёмом корпуса.

Данные вывод означает, что составленные корпуса не являются насыщенными в рамках определенных рассматриваемых моделей s-грамм. То есть увеличение объема исходного материала ведет к пропорциональному

пополнению словарей новыми элементами. Вопрос насыщенности корпусов тесно связан с важной информационной характеристикой – покрытием корпусов или словарей.

3. Результаты исследования моделей покрытия и экспериментальной проверки позволяют заключить следующее:
   - Покрытие словарей из газетно-публицистического корпуса, оказалось выше покрытия словарей из художественного корпуса того же объема (по результатам экспериментов покрытие 10-грамм – больше в 2 раза). Данный факт объясняется меньшим лексическим разнообразием публицистических текстов и наличием часто повторяющихся оборотов по сравнению с литературными текстами.
   - Увеличение материала, как и ожидалось, ведет к возрастанию покрытия. Например, согласно обеим моделям для английского языка увеличение корпуса в 10 раз (с 10^7 до 10^8 символов) ведет к практически двукратному росту покрытия.
   - На одинаковом объеме исходных корпусов для английского языка удается достичь большего покрытия словарей s-грамм. Например, при исследовании словарей из корпуса 10^7 символов покрытие 10-грамм для русского языка составило 20%, а для английского – 33%.
   - Точность оценок предложенных моделей покрытия увеличивается с ростом s.
4. По результатам исследования предельный уровень энтропии составил 0,78-0,91 бит на символ для русского языка и 0,55-0,87 бит на символ для английского языка. Отметим, что в опытах по исследованию энтропии текстов (русской прозы), проводившихся Колмогоровым, оценка предельного уровня энтропии также была получена, начиная с H_{50} [17]. Сравнение оценок предельного уровня энтропии в данном диссертационном исследовании и ранее полученных результатов Пиотровского и Шеннона приведены ниже (см. Таблица 2).

Таблица 2 – Известные оценки предельного уровня энтропии H_0

Оценка	Русский язык: публицистика и научно-деловая речь	Английский язык
известные результаты (Пиотровский и Шеннон соответственно)	0,59-0,83 бит/симв.	0,6-1,3 бит/симв.
полученная в данном исследовании	0,59-0,89 бит/симв.	0,55-0,87 бит/симв.

Источник: результаты, полученные автором, а также результаты Р. Г. Пиотровского (1968) и К. Шеннона (1951)

В **третьей главе** разрабатывается и исследуется метод восстановления участков входных сообщений с использованием теоретико-информационного подхода, в том числе теоретические и экспериментальные методы оценки доли восстановленных участков сообщений. Проведен анализ методов полного восстановления исходного сообщения, выяснены условия невозможности полного восстановления сообщения. Рассмотрена задача восстановления отдельных участков (s-грамм) сообщения в случае невозможности полного восстановления текста. Введена модель канала связи, на выходе которого задаются теоретико-вероятностные распределения появления информации о символах входного сообщения. В рамках модели канала связи разработан метод восстановления отдельных s-грамм сообщения, основанный на поиске подходящих s-грамм и восстановлению данных s-грамм по заранее сформированному словарю в рамках границ допустимой многозначности. Оценена трудоемкость данного метода.

Для проведения экспериментального исследования метода разработана программная реализация алгоритма восстановления отдельных участков сообщения на языке C++. Введены теоретико-вероятностные модели появления значений символов на выходе канала связи: смоделированы дискретное равномерное распределение, полиномиальное распределение и распределение, возникающее в случае многократного применения ключа в поточных преобразованиях. Исследование проводится в рамках различных теоретико-вероятностных моделей появления выходных наборов. Проведено экспериментальное исследование по восстановлению подходящих s-грамм и

оценена средняя доля s-грамм, которую удается восстановить на выходе канала связи. Экспериментальное исследование имеет следующую структуру:

1. Исследование восстановления s-грамм в рамках заданных вероятностных распределений появления значений символов на выходе канала связи:
    a. восстановление s-грамм сообщения, которое является частью исходного корпуса словарей (т. е. истинная s-грамма всегда присутствует в словарях);
    b. восстановление s-грамм сообщения, которое **не является** частью исходного корпуса словарей (т. е. истинная s-грамма может отсутствовать в словарях из-за неполного покрытия).
2. Исследование восстановления s-грамм в случаях многократного применения ключа в поточных преобразованиях:
    a. вероятность совпадения символов неизвестного сообщения с одним из символов известного*;*
    b. восстановление s-грамм сообщения, которое является частью исходного корпуса словарей (т. е. истинная s-грамма всегда присутствует в словарях).

Помимо экспериментального исследования, получены теоретические оценки доли восстановленных s-грамм, которая зависит от вероятности появления подходящей s-граммы в сообщении при фиксированной критической границе отбора, и вероятности, что восстановление подходящей s-граммы не превысит границу допустимой многозначности при фиксированных параметрах. Получена теоретическая оценка вероятности появления подходящих s-грамм в сообщении в рамках заданной теоретико-вероятностной модели на выходе канала связи. Построена модель оценки с помощью нормального распределения. Получена теоретическая оценка вероятности появления подходящих s-грамм в случае многократного использования ключа в поточных преобразованиях. Показано, что такую вероятность можно оценить сверху и снизу с помощью биномиального

распределения, оценки параметра которого получены при экспериментальном исследовании в п. 2а. Получена теоретическая оценка вероятности появления допустимых вариантов s-граммы при восстановлении. Построена модель оценки с помощью гипергеометрического распределения, проведен асимптотический анализ полученных формул. Проведена экспериментальная проверка предложенных моделей аппроксимации вероятностных распределений. Выполнены численные расчеты данных моделей.

Выводы по главе 3

По результатам исследования, описанного в данной главе, можно сделать нижеследующие выводы.

1. Исследование в рамках заданной теоретико-вероятностной модели без учета покрытия словаря:
   - Для русского языка рассмотрение 10-грамм позволяет восстановить не более 0,05 доли s-грамм сообщения. При рассмотрении 15-грамм возможно восстановить около 0,50 доли s-грамм исходного сообщения, увеличивая критическую границу отбора подходящих участков (16 и более символов). При рассмотрении 20- и 25-грамм увеличение критической границы приводит к соответствующему росту доли восстановленных s-грамм до 1. Стоит учитывать, что покрытие экспериментальных словарей 20 и 25-грамм не превышало 4% и 3% соответственно.
   - Для английского языка рассмотрение 10-грамм привело к восстановлению не более 0,05 доли s-грамм. При рассмотрении 15-грамм удалось восстановить чуть менее 0,25 доли s-грамм. При рассмотрении 20- и 25-грамм средняя доля растет до 1 (с увеличением критической границы).
2. Исследование в рамках заданной теоретико-вероятностной модели с учетом возможной неполноты покрытия словаря:
   - Для русского языка рассмотрение 15-грамм позволило восстановить наибольшую долю (0,17) s-грамм сообщения по сравнению с другими длинами s-грамм. При рассмотрении других длин s-грамм восстановленная доля не превышала 0,05.
   - Для английского языка рассмотрение 15-грамм и 20-грамм позволило восстановить наибольшую долю (0,16 и 0,17 соответственно) s-грамм сообщения по сравнению с другими длинами s-грамм. При рассмотрении других длин s-грамм восстановленная доля не превышала 0,03.

3. По результатам проведенных экспериментов можно сделать следующие выводы:
    - Вероятность совпадения символов текстов сообщений для английского языка несколько выше, чем для русского языка.
    - При двукратном применении ключа удалось восстановить менее 0,0001 доли 20-грамм и 25-грамм на русском языке (без учета покрытия словарей); при трехкратном применении ключа удалось восстановить около 0,05 25-грамм; при четырехкратном – 0,22, при пятикратном – 0,66 (без учета покрытия словарей).

Выводы по диссертации

1. Доля при восстановлении отдельных s-грамм сообщения при определённых параметрах может достигать более 0,9, однако данный результат приведен без учета покрытия используемых словарей. Тем не менее, в ряде практических случаев возможно ограничить множество входных сообщений узким классом, в котором тексты стандартны по своей структуре и ограничиваются несколькими повторяющимися темами (например, в ранее исследовались теоретико-информационные свойства языка наземного управления полетами). В таком случае можно добиться покрытия словарей, близкого в полному в рассматриваемых условиях.
2. В более общих случаях, когда на содержание и структуру текста не накладывается значительных ограничений, добиться практически приемлемого покрытия сложно, особенно для 20- и 25-грамм. Наложение некоторых ограничений на стиль и тематику текстов позволяет увеличить покрытие словарей (например, покрытие 10-грамм художественных текстов в 2 раза меньше покрытия газетно-публицистических текстов, ограниченных политической тематикой).
3. Наибольшего покрытия удалось достичь при рассмотрении 10-грамм (40% для русского и 54% для английского языков), даже не накладывая стилистических ограничений на тексты исходного корпуса. Но с практической точки зрения, как показало настоящее исследование, рассмотрение 10-грамм позволяет восстановить лишь незначительную долю s-грамм из-за высокой многозначности.
4. С учетом экспоненциального уменьшения покрытия словаря s-грамм с ростом s, наибольшую долю восстановленных s-грамм удалось получить при рассмотрении 15-грамм для русского языка (до 0,17) и 20- и 25-грамм символов для английского (0,16 и 0,17 соответственно).
5. Выбор критической границы отбора подходящих s-грамм зависит от языка исходного текста (например, мощности его алфавита). Например,

в рамках заданной теоретико-вероятностной модели появления вариантов знаков выбор критической границы отбора s-грамм, равной 14 символам, позволил восстановить около 0,5 доли s-грамм сообщения для русского языка и более 0,8 – для английского языка (без учета покрытия словарей).

6. Ниже (см. Таблица 3) приведены результаты восстановления 15-грамм сообщения, не являющегося частью исходного корпуса словарей в рамках заданной теоретико-вероятностной модели (то есть с учетом влияния неполноты покрытия словарей).

   Таблица 3 – Доля восстановленных 15-грамм с учетом возможной неполноты словарей

Критическая граница	*Доля 15-грамм*	
	Русский язык	Английский язык
12 символов	0,03	0,13
16 символов	0,17	0,16

Источник: рассчитано автором.

7. Разница в теоретико-информационных свойствах языков и мощностных характеристиках исходных алфавитов обуславливает разную долю восстановленных s-грамм при идентичных параметрах.

Результаты, выносимые на защиту

1. Алгоритмы нормализации текстовых корпусов и построения словарей s-грамм, позволяющие сформировать прикладную базу и подготовить эмпирический материал для проведения анализа характеристик текстов на естественном языке.
2. Модели покрытия словарей s-грамм, позволяющие получить численные оценки информационных характеристик s-грамм на естественном языке.
3. Метод оценки энтропии коротких s-грамм, основанный на мощностных характеристиках словарей s-грамм. Данный метод подразумевает подход, основанный на статистических характеристиках текстов, без привлечения вероятностной модели для построения оценок.
4. Оценки допустимой многозначности s-грамм, возникающей при восстановлении текстов на русском и английском языках. Данный результат позволяет учесть влияние языковой вариативности на алгоритмы восстановления текста.
5. Метод восстановления участков входных сообщений, основанный на применении теоретико-информационного подхода. Данный метод позволяет осуществлять поиск и восстановление подходящих s-грамм сообщения и применим в случае невозможности полного восстановления сообщения.

Заключение

В рамках диссертационной работы:

1. Разработаны алгоритмы нормализации текстовых корпусов и построения словарей s-грамм, позволяющие сформировать прикладную базу для проведения анализа характеристик текстов на естественном языке на основе теоретико-информационного подхода.
2. Предложены и исследованы математические модели покрытия словарей s-грамм, основанные на информационных характеристиках текстов на естественном языке. Данные модели позволяют оценить покрытие экспериментально составляемых словарей и соответствующую ошибку I рода (потерять истинный вариант), возникающую при восстановлении s-грамм по словарю.
3. Разработан метод оценки энтропии коротких s-грамм, основанный на мощностных характеристиках словарей s-грамм.
4. Проведены исследования границ допустимой многозначности восстановления s-грамм сообщения.
5. Разработан метод восстановления участков входных сообщений с использованием теоретико-информационного подхода.
6. Получены оценки доли восстановленных s-грамм сообщения с использованием словарей для русского и английского языков в рамках определенной теоретико-вероятностной модели появления значений символов на выходе канала и в случаях многократного использования ключа поточного преобразования.
7. Получены теоретические оценки доли восстановленных s-грамм сообщения в рамках различных теоретико-вероятностных моделей.

Публикации и апробация работы

Публикации автора в Scopus/WoS:

1. Malashina A. Possibility of Recovering Message Segments Based on Side Information about Original Characters // Doklady Mathematics. 2024. Vol. 108. No. Suppl 2. (**Scopus Q2**, WoS Q3, MathSciNet) *(перевод с русского)*
2. Малашина А. Г. О возможности восстановления отрезков сообщения по информации о значениях исходных символов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514. № 2. С. 138-149. (**Список В**)
3. Malashina A. The Combinatorial Analysis of n-Gram Dictionaries, Coverage and Information Entropy based on the Web Corpus of English // Baltic Journal of Modern Computing, 2021, Т.9, №3, С. 363-376. (**Scopus Q4,** WoS Q4, Список С)
4. Малашина А. Г. Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов / Малашина А. Г., Лось А. Б. // Чебышевский сборник, 2022, Т.23, №2, С. 151-160. (**Scopus Q3,** Список С)

Публикации в иных изданиях:

5. Малашина А. Г. Разработка инструментальных средств для исследования информационных характеристик естественного языка // Промышленные АСУ и контроллеры. 2021. № 2. С. 9-15. *(ВАК К2)*
6. Малашина А. Г. Модификация одного алгоритма восстановления текстовых сообщений и математическая модель распределения числа осмысленных текстов // Электронные средства и системы управления: материалы докладов XVI Международной научно-практической конференции (18–20 ноября 2020 г.): в ч. 2. Томск: В-спектр, 2020. С. 85-88. *(РИНЦ)*
7. Малашина А. Г. Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов / Малашина А. Г., Лось А. Б. // Алгебра, теория чисел и дискретная геометрия: современные

проблемы, приложения и проблемы истории. Тула: ТГПУ, 2020. С. 177-181. *(РИНЦ)*

Доклады на конференциях:

1. Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов им. Е.В. Арменского, Россия, Москва, 27.02.2023 – 07.03.2023, доклад: «Исследование энтропии текстов на естественном языке методом составления словарей n-грамм».
2. Entropy 2021: The Scientific Tool of the 21st Century, Португалия, Порту, 05.05.2021 – 07.05.2021, доклад: «Entropy analysis of n-grams and estimation of the number of meaningful language texts».
3. Всероссийский конкурс-конференция студентов и аспирантов по информационной безопасности «SIBINFO-2021», Россия, Томск, 22.04.2021, доклад: «Математическая модель алгоритма восстановления отдельных частей текстового сообщения».
4. XXIII научно-практическая конференция «РусКрипто’2021», Россия, Солнечногорск, 23.03.2021 – 26.03.2021, доклад: «Алгоритм восстановления отдельных частей текстовых сообщений по информации о возможных вариантах его знаков».
5. Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов им. Е.В. Арменского, Россия, Москва, 10.03.2021 – 17.03.2021, доклад: «Предельные распределения, возникающие в задаче восстановления отдельных отрезков текстового сообщения».
6. XVI Международная научно-практическая конференция «Электронные средства и системы управления», Россия, Томск, 18.11.2020 – 20.11.2020, доклад: «Модификация одного алгоритма восстановления текстовых сообщений и математическая модель распределения числа осмысленных текстов».
7. XVIII Международная научная конференция «Алгебра, теория чисел и дискретная геометрия: современные проблемы, приложения и проблемы

истории», посвященная 100-летию со дня рождения профессоров Б. М. Бредихина, В. И. Нечаева и С. Б. Стечкина, Россия, Тула, 23.09.2020 – 26.09.2020, доклад: «Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов».

8. Всероссийский конкурс-конференция студентов и аспирантов по информационной безопасности «SIBINFO-2020», Россия, Томск, 16.04.2020, доклад: «Модификация одного алгоритма восстановления текстовых сообщений и математическая модель распределения числа осмысленных текстов».
9. Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов им. Е.В. Арменского, Россия, Москва, 25.02.2020 – 04.03.2020, доклад: «Статистический анализ языковых моделей русского языка на основе текстового новостного корпуса».

Доклады на семинарах научного руководителя МИЭМ НИУ ВШЭ и кафедры компьютерной безопасности НИУ ВШЭ:

1. Семинар от 23 апреля 2024 г., доклад: «Исследование информационных характеристик естественных языков в связи с разработкой методов оценки защищенных информационных систем».
2. Семинар от 18 апреля 2023 г., доклад: «Исследование информационных характеристик естественных языков в связи с разработкой методов оценки защищенных информационных систем».
3. Семинар от 25 ноября 2021 г., доклад: «Исследование информационных характеристик естественных языков в связи с разработкой методов оценки защищенных информационных систем».

Результаты интеллектуальной деятельности:

1. Свидетельство №2022662474 о регистрации программы для ЭВМ «Программа для восстановления отельных участков сообщения по информации о возможных символах его знаков», 2022 г.

2. Свидетельство № 2020665906 о регистрации программы для ЭВМ «Программа для создания словарей n-грамм и вычисления их информационных характеристик», 2020 г.

Проекты:

Результаты диссертационного исследования были также использованы при выполнении проектов под руководством автора:

1. Студенческий проект МИЭМ №338, «Исследование информационных характеристик естественных языков[1]», 2020-2021 гг.
2. Проект «Исследование информационных характеристик естественных языков» в рамках «Ярмарки проектов», 2020 г.

[1] Результаты проекта доступны в публикации *Нагаева И. Э., Савченкова Д. М. Исследование информационных характеристик художественных текстов и их переводов // Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени ЕВ Арменского. – 2021. – С. 241-244.*

Список литературы, использованной в резюме

1. Babash A. V. Attacks on the Random Gamming Code // Mathematics and Mathematical Modeling, 2020, no. 6, pp. 35–38.
2. Brown P., Della Pietra V., Della Pietra S., Lai J., Mercer R. An estimate of an upper bound for the entropy of English // Computational Linguistics, 1992, vol. 18(1), pp. 31–40.
3. Calin O. Statistics and machine learning experiments in English and Romanian poetry // Sci 2(4), 2020, https://doi.org/10.3390/sci2040092
4. Florencio D., Herley C. A large-scale study of web password habits // Proceedings of the 16th international conference on World Wide Web, Association for Computing Machinery: New York, USA, 2007, pp. 657–666
5. Kolmogorov A. Three approaches to the definition of the notion of amount of information // Shiryayev, A.N. (eds) Selected Works of A. N. Kolmogorov, 1993, pp. 184-193.
6. Kontoyiannis I., Algoet P. H., Suhov Y. M., Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text // IEEE Trans. Inf. Theory 1998, 44, pp. 1319–1327.
7. Shannon C. E. A mathematical theory of communication // The Bell system technical journal, 1948, 27(3), pp. 379–423.
8. Shannon C. E. Prediction and entropy of printed English // Bell system technical journal, 1951, 30(1), pp. 50–64.
9. Teahan W., Cleary J. The entropy of English using ppm-based models // Proceedings of Data Compression Conference - DCC’96, IEEE, Snowbird, UT, USA, 1996, pp. 53–62.
10. Деундяк В. М., Пашкова Е. А. Математические вопросы криптоанализа шифра гаммирования в случае серии ошибок оператора типа пропуск-повтор [Электронный ресурс], 2005. URL: http://scholar.googleusercontent.com/scholar?q=cache:KJEuFdgLDG0J:scholar.google.com/+Математические+вопросы+криптоанализа+шифра+гаммир

ования+в+случае+серии+ошибок+оператора+типа+пропуск-повтор+&hl=ru&as_sdt=0,5 (дата обращения: 13.02.2024)

11. Диссертационный совет по инженерным наукам и прикладной математике. Паспорт области науки «Инженерные науки и прикладная математика», утвержденный ученым советом НИУ ВШЭ от 2 февраля 2018 г. URL: https://www.hse.ru/mirror/pubs/share/335373657
12. Лось А. Б., Нестеренко А. Ю., Рожков М. И. Криптографические методы защиты информации // Издательство Юрай, 2016, с. 56-65.
13. Малашина А. Г. О возможности восстановления отрезков сообщения по информации о значениях исходных символов // Доклады Российской академии наук. Математика, информатика, процессы управления, 2023, т. 514. № 2, с. 138-149.
14. Пиотровский Р. Г. Информационные измерения языка // М: Издательство «Наука», 1968.
15. Проскурин Г. В. Принципы и методы защиты информации // МИЭМ: Москва, 1997.
16. Шеннон К. Работы по теории информации и кибернетике // М: Издательство иностранной литературы, 1963, с. 669- 687.
17. Яглом А. М., Яглом И. М. Вероятность и информация. Издание третье, переработанное и дополненное. М.: Наука. 1973, с. 236-312.
18. Нагаева И. Э., Савченкова Д. М. Исследование информационных характеристик художественных текстов и их переводов // Межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского. 2021. С. 241-244.