

Министерство высшего образования и науки Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
САЕ «Институт человека цифровой эпохи»
Автономная магистерская программа «Компьютерная и когнитивная лингвистика»

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК

Руководитель ООП
д-р филол. наук, профессор
Резанова З.И. Резанова
« 19 » июня 2021 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

АНАЛИЗ РЕЗУЛЬТАТОВ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ НОВОСТНЫХ ТЕКСТОВ

по основной образовательной программе подготовки магистров
направление подготовки 45.04.03 – Фундаментальная и прикладная лингвистика

Белогорская Дарья Викторовна

Руководитель ВКР,
д-р. филол. наук, профессор
Резанова З.И. Резанова
подпись
« 19 » июня 2021г.

Консультант,
науч. сотр. Лаборатории
когнитивных
исследований языка
Степаненко А.А. Степаненко
подпись
« 19 » июня 2021г.

Автор работы
студент группы № 131983
Белогорская Д.В. Белогорская
подпись

Министерство высшего образования и науки Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
САЕ «Институт человека цифровой эпохи»
Автономная магистерская программа «Компьютерная и когнитивная лингвистика»

УТВЕРЖДАЮ

Руководитель АМП
д-р филол. наук, профессор
Резанова 3.И. Резанова
«15» сентября 2020 г.

ЗАДАНИЕ

по подготовке магистерской диссертации
студенту Белогорской Дарье Викторовне, группы 131983

1. Тема магистерской диссертации: «Анализ результатов автоматического реферирования новостных текстов».
2. Срок сдачи студентом выполненной магистерской диссертации:
а) на кафедре: 26.06.2021 г.
б) в ГЭК: 26.06.2021 г.
3. Исходные данные к работе:

Объект исследования: автоматическое реферирование новостных текстов.

Предмет исследования: алгоритмы реферирования новостных текстов.

Цель: разработать систему оценки эффективности алгоритмов реферирования, примененных к материалу новостных текстов.

Задачи:

1. Изучить существующие методы реферирования.
2. Изучить теоретико-методологические основы автоматического реферирования текстов.
3. Ознакомиться с существующими алгоритмами автоматического реферирования.
4. Ознакомиться с жанровой структурой новостных текстов.
5. Найти, отобрать, извлечь и сформировать корпус новостных текстов для анализа.
6. Разработать критерии оценки качества рефератов.
7. Применить готовые алгоритмы реферирования к корпусу новостных текстов.
8. Проанализировать результаты работы разных методов реферирования с лингвистической точки зрения.
9. Сравнить результаты реферирования, проведенного на основе применения разных алгоритмов.

Используемые методы: метод автоматического анализа текста, включающий следующие приемы: сбор данных, предобработка текста, применение программного

кода для выполнения автоматического реферирования; метод филологического анализа текста: оценка; метод сплошной выборки.

4. Краткое содержание работы:

Магистерская диссертация состоит из введения, двух глав, заключения, списка использованной литературы и приложений.

Во введении обосновывается актуальность исследования, формулируются объект, предмет, цель и решаемые задачи, а также характеризуется теоретическая основа исследования.

В первой главе рассмотрены такие ключевые пункты, как зарождение и становление автоматического реферирования, проблемы и методы автоматического реферирования, описаны подходы и этапы к автоматическому реферированию, а также раскрыты особенности новостных текстов как объектов реферирования.

Вторая глава посвящена анализу результатов автоматического реферирования новостных текстов.

В заключении описываются реализованные этапы исследования и определяются перспективы его дальнейшего развития.

В Приложении А представлены иллюстрации программного кода на языке программирования Python.

В Приложении Б представлены примеры хороших текстов обоих алгоритмов.

5. Предприятие/организация, по заданию которого выполняется работа: Томский государственный университет.

6. Дата выдачи задания: «15» сентября 2020 г.

Руководитель магистерской диссертации,
д-р филол. наук, профессор

Задание принял к исполнению

 З.И. Резанова

 Д.В. Белогорская

Аннотация

к магистерской диссертации на тему Анализ результатов автоматического реферирования новостных текстов».

Данная исследовательская работа проводится в сфере информационных технологий и направлена на анализ результатов работы алгоритмов реферирования новостных текстов.

Для достижения цели – разработать систему оценки эффективности алгоритмов реферирования, примененных к материалу новостных текстов – был изучен теоретический материал по теме «ручного» и автоматического реферирования, а также по теме медиалингвистики, была собрана и проанализирована выборка из 100 новостных текстов, в которую вошли новостные текста Информационного агентства «ТАСС».

В ходе практической работы были разработаны критерии оценки качества автоматического реферирования, которые в дальнейшем применялись для анализа выборки. Для того, чтобы получить анализируемые выборки, были отобраны два алгоритма автоматического реферирования, через которые были пропущены новостные текста.

Результатом исследования является таблица с соотношением критериев оценки качества и количеством ошибок по каждому из критериев в двух анализируемых алгоритмах.

В первой главе рассмотрены такие ключевые пункты, как зарождение и становление автоматического реферирования, проблемы и методы автоматического реферирования, описаны подходы и этапы к автоматическому реферированию, а также раскрыты особенности новостных текстов как объектов реферирования.

Во второй главе рассмотрены алгоритмы автоматического реферирования, разработаны критерии оценки качества автоматического реферирования, описаны примеры анализа автоматически сгенерированных рефератов на основе разработанных критериев, представлены и описаны результаты анализа автоматически сгенерированных рефератов на основе разработанных критериев.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	4
ГЛАВА 1	8
Теоретико-методологические основы «ручного» и автоматического реферирования текстов	8
1.1 Реферирование как способ извлечения и представления основного содержания текста	8
1.1.1 Становление реферирования	8
1.1.2 Определение понятий «реферирование» и «реферат»	9
1.1.3 Методы реферирования	10
1.1.4 Виды рефератов	13
1.1.5 Сущность и методы компрессии материала первоисточника	15
1.2 Проблемы и методы автоматического реферирования текстов	17
1.2.1 Зарождение и становление идеи автоматического реферирования текста	17
1.2.2 Автоматическое реферирование текста	19
1.2.2.1 Переход от «ручного» к автоматическому реферированию	19
1.2.2.2 Определение понятия «автоматическое реферирование»	20
1.2.2.3 Преимущества автоматического реферирования	21
1.2.2.4 Этапы и подходы к автоматическому реферированию	21
1.2.2.5 Проблемы автоматического реферирования, проверка качества	23
1.2.2.6 Варианты автоматического реферирования в зависимости от типов текстов	25
1.3 Особенности новостных текстов как объектов реферирования	25
1.3.1 Значимость массмедиа	25
1.3.2 Язык массмедиа	26
1.3.3. Предмет медиалингвистики	27
1.3.3.1 Структура и содержание понятия «медiateкст»	28
1.3.3.2 Методы изучения медиатекстов	30
1.3.3.3 Типы медиатекстов: новости	31
1.3.3.4 Лингвомедийные свойства новостных текстов. Общая характеристика новостных текстов. Понятие новостной ценности	31
1.3.3.5 Особенности формата новостных текстов в средствах массовой информации в сети Интернет	34
Выводы по первой главе	36
ГЛАВА 2	37
Лингвистический анализ результатов автоматического реферирования новостных текстов	37
2.1 Алгоритмы автоматического реферирования	37
2.2 Критерии оценки качества реферата	40
2.2.1 Метрики BLEU и ROUGE для оценки качества автоматического реферирования	40
2.2.2 Критерии оценки «ручного» реферирования	40
2.2.3 Анализ автоматически сгенерированных рефератов	44

2.2.3.1 Пример анализа автоматически сгенерированных рефератов на основе разработанных критериев	44
2.2.3.3 Результаты анализа автоматически сгенерированных рефератов на основе разработанных критериев.....	51
Выводы по второй главе	59
ЗАКЛЮЧЕНИЕ.....	61
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ.....	63
ПРИЛОЖЕНИЕ А.....	70
ПРИЛОЖЕНИЕ Б.....	72

ВВЕДЕНИЕ

Данная исследовательская работа проводится в сфере информационных технологий и направлена на анализ результатов работы алгоритмов реферирования новостных текстов.

В настоящее время все мы являемся очевидцами того, как быстро информационные потоки заполняют интернет-пространство. Стремительный рост объема информации именуется «информационным взрывом». Вместе с потоками информации растет и информационная перенасыщенность людей, которая приводит к неспособности адекватно воспринимать и перерабатывать необходимый объем поступающей информации. Большая часть информации представлена в текстовом формате, поэтому для решения проблемы «информационного взрыва» необходимо участие компьютерных лингвистов. С помощью методов, применяемых в компьютерной лингвистике, можно проводить автоматический анализ текстовых данных на естественном языке. Процесс автоматического реферирования текстовой информации экономит ресурсы, так как превосходит по скорости ручной труд человека и сокращает материальные затраты на содержание штата сотрудников.

Автоматическое реферирование текстов можно осуществлять разными способами. Например, существуют подходы, основанные на графах. Представителем такого подхода является TextRank. Для автоматического реферирования также используется алгоритм LSA (Latent Semantic Analysis), название алгоритма говорит само за себя, так как реферирование текста основывается на скрытом семантическом анализе. Реферирование текстов осуществляется и посредством алгоритма k-means, который основывается на кластеризации. Каждый из способов имеет свои недостатки и преимущества, одни употребляются чаще других.

В данной исследовательской работе будет уделено особое внимание автоматическому реферированию на основе статистических методов.

Актуальность настоящей работы обуславливается потребностью в разработке качественного и работающего алгоритма автоматического реферирования новостных текстов. Данная потребность вызвана «информационным взрывом», который сейчас претерпевает современное общество. Кроме того, в использовании алгоритмов автоматического реферирования текстов заинтересованы различные новостные агентства и порталы. Повышенное внимание к проблеме информационной перенасыщенности говорит о том, что данная проблема весьма актуальна. Более того, существующие алгоритмы автоматического реферирования текстов применяются в основном на англоязычных текстах, поэтому русскоязычное общество нуждается в адаптации данных алгоритмов или

в создании новых, которые будут исправно работать на текстах, написанных на русском языке.

В настоящее время в научном сообществе часто поднимается вопрос автоматического реферирования текстов, что доказывается активным обсуждением данной темы на конференциях по компьютерной лингвистике. Представителем одной из таких конференций является «Диалог» — крупнейшая в России международная научная конференция по компьютерной лингвистике. Начиная с конца прошлого века, количество исследований в данной области увеличилось в разы. Кроме того, проекты по автоматическому реферированию текстов активно финансируются.

Таким образом, высокая социальная значимость, нарастающий интерес со стороны заинтересованных в автоматическом реферировании лиц (новостные агентства и компании), проблема адаптации методов автоматического реферирования англоязычных текстов к русскому языку, а также необходимость в разработке системы оценки их результативности доказывают актуальность обращения к реферированию новостных текстов.

Объектом исследования является автоматическое реферирование новостных текстов.

Предметом исследования выступают алгоритмы реферирования новостных текстов.

Цель — разработать систему оценки эффективности алгоритмов реферирования, примененных к материалу новостных текстов.

Поставленная цель определила следующие основные **задачи** работы:

1. Изучить существующие методы реферирования.
2. Изучить теоретико-методологические основы автоматического реферирования текстов.
3. Ознакомиться с существующими алгоритмами автоматического реферирования.
4. Ознакомиться с жанровой структурой новостных текстов.
5. Найти, отобрать, извлечь и сформировать корпус новостных текстов для анализа.
6. Разработать критерии оценки качества рефератов.
7. Применить готовые алгоритмы реферирования к корпусу новостных текстов.
8. Проанализировать результаты работы разных методов реферирования с лингвистической точки зрения.
9. Сравнить результаты реферирования, проведенного на основе применения разных алгоритмов.

Материалом исследования послужили текстовые данные информационного агентства «ТАСС» из рубрики по происшествиям [16]. Всего методом сплошной выборки было отобрано 100 новостных текстов из рубрики «Происшествия».

С учетом поставленной цели в данном исследовании применяются следующие **методы**: метод автоматического анализа текста, включающий следующие приемы: сбор данных, предобработка текста, применение программного кода для выполнения автоматического реферирования; метод филологического анализа текста: оценка; метод сплошной выборки.

Метод сплошной выборки применялся на этапе сбора текстовых материалов для анализа. Тексты отбирались по мере их встречаемости в рубрике «Происшествия».

Для подготовки текстовой выборки к анализу применялись методы предварительной обработки текстов. К предобработанной текстовой выборке осуществлялось применение программного кода с целью получить рефераты.

Метод филологического анализа, в частности, оценка, применялся на этапе анализа полученных рефератов.

Теоретическая основа исследования.

В области автоматического реферирования мы опирались на работы Х. Луна, Х. Торреса-Морено, Г.П. Эдмундсона, Р.И. Уиллиса, К. Спарк Джонс, Д.Б. Клевеленда, К. МакКеона и Б. Дорра.

В области медиалингвистики и жанровой структуры новостных текстов мы опирались на работы Т.Г. Добросклонской, Т. Ван Дейка, А.Н. Богомолова, В.Г. Костомарова, Б.В. Кривенко, О. А. Ксендзенко, М.М. Назарова, А.А. Гречихина, И.Г. Здорова, В.И. Соловьева, А.И. Жолкова, Л.П. Маркушевой, Ю.А. Цапаевой, Д.И. Блюменау, Н.М. Нестеровой, Н.А. Герте, В.А. Вейзе и др.

Структура работы определяется целями и задачами. Магистерская диссертация состоит из введения, двух глав, заключения, списка использованной литературы и приложений.

Во введении обосновывается актуальность исследования, формулируются объект, предмет, цель и решаемые задачи, а также характеризуется теоретическая основа исследования.

В первой главе рассмотрены такие ключевые пункты, как зарождение и становление автоматического реферирования, проблемы и методы автоматического реферирования, описаны подходы и этапы к автоматическому реферированию, а также раскрыты особенности новостных текстов как объектов реферирования.

Вторая глава посвящена анализу результатов автоматического реферирования новостных текстов.

В заключении описываются реализованные этапы исследования и определяются перспективы его дальнейшего развития.

В Приложении А представлены иллюстрации программного кода на языке программирования Python.

В Приложении Б представлены примеры хороших текстов обоих алгоритмов.

Результаты исследования были апробированы на VII (XXI) Международной научно-практической конференции «Актуальные проблемы лингвистики и литературоведения» и V Международному научно-исследовательском конкурсе «Конкурс молодых ученых».

ГЛАВА 1

Теоретико-методологические основы «ручного» и автоматического реферирования текстов

1.1 Реферирование как способ извлечения и представления основного содержания текста

1.1.1 Становление реферирования

Известное нам со школьной скамьи понятие реферирования как процесса извлечения из текста наиболее существенной информации и ее последующее представление в рамках вторичного текста стало неотъемлемой частью современной жизни, так как реферирование можно назвать одним из основных способов разрешения перегруженности информационного потока. Несмотря на то, что реферирование активно применяется к текстам в наше время, оно имеет многовековую историю, которая, как отмечают Н.И. Нестерова и Н.А. Герте, корнями уходит к Шумерской цивилизации. Во времена Шумера писцы на глиняных табличках создавали краткие записи, свидетельствующие о столь раннем зарождении реферирования. Далее, углубляясь в историю, можно отследить использование реферирования в эпоху распада Римской империи, когда на пергаментях делались краткие тезисы и описания различного рода документов. Предположительно именно в этот период времени возник термин *abstractus* (с лат. – абстрактный). Впервые реферативные журналы в Европе появились в XVII-XVIII вв. Такие журналы предназначались для информирования ученых о научных открытиях и достижениях. В России первым реферативным центром стала Петербургская академия наук, а первыми реферативными журналами можно считать следующие: «Краткое описание комментариев академии наук» и «Содержание ученых рассуждений» [26].

М.В. Ломоносов считается первым российским теоретиком реферирования. Он не только говорил о необходимости и пользе реферирования, но и сформулировал основные правила реферирования для референтов-журналистов: «Всякий, кто берет на себя труд осведомлять публику о том, что содержится в новых сочинениях, должен прежде всего взвесить свои силы. Ведь он затевает трудную и очень сложную работу, при которой приходится докладывать не об обыкновенных вещах и не просто об общих местах, но схватывать то новое и существенное, что заключается в произведениях, создаваемых часто величайшими людьми. Высказывать при этом неточные и безвкусные суждения – значит сделать себя предметом презрения и насмешки; это значит уподобиться карлику, который хотел бы поднять горы» [9]. Термин «реферат» впервые появился в России в «Настольном словаре для справок по всем отраслям знаний» Ф. Г. Толля. В словаре данный термин имел следующее обозначение: «отношение, деловая записка, изложение дела вкратце» [19].

В 1952 г. в России создается Всероссийский институт научной и технической информации РАН (ранее Всесоюзный институт научной и технической информации АН СССР). Создание данного института стало важнейшим шагом в развитии реферирования в России, так как на базе института стали выпускаться тематические реферативные журналы ВИНТИ.

1.1.2 Определение понятий «реферирование» и «реферат»

Данная исследовательская работа проводится в области автоматического реферирования, но прежде чем более детально говорить об автоматическом реферировании текстов, следует разобраться в реферировании, которое осуществляется человеком. Погружение в тему реферирования мы начнем с определения понятий «реферирование» и «реферат».

В ГОСТе понятие реферирования определяется так: «Реферирование: процесс составления реферата» [7]. Данное определение не раскрывает в полной мере понятие реферирования, так как в качестве определения дается достаточно краткая формулировка, которая скорее является отсылкой к понятию «реферат», нежели полноценным определением понятия реферирования. Так, например, определение понятия реферирования А.В. Бастрикова, которое автор дает в своей работе, имеет более детальную развернутую структуру, однако в самом определении упоминается, что реферирование приравнивается к составлению реферата, на что указывает определение из ГОСТа. По А.В. Бастрикову реферирование, или составление реферата (лат. *refferere* - докладывать, сообщать), – это краткое изложение содержания оригинала-первоисточника [30]. По сравнению с первым определением из ГОСТа данное определение в большей степени раскрывает суть реферирования и дает краткое объяснение процесса создания реферата.

Выше были рассмотрены варианты определений понятия «реферирование» на отечественном материале. Далее понятие «реферирование» будет рассмотрено в рамках компьютерной лингвистики на материале зарубежных авторов, так как в большинстве своем работы в области компьютерной лингвистики пишутся на английском языке. Так, в компьютерной лингвистике выделяют разные варианты определения понятия реферирования, в данной исследовательской работе будут рассмотрены некоторые из них. Т.А. Ван Дейк в своей книге пишет, что реферирование – это указание и прогнозирование структуры и содержания текста [83]. Х. М. Торрес-Морено в своей работе ссылается на определение, которое дает Д.Б. Клевеленд: «Реферирование суммирует основное содержание конкретной текстовой записи и является истинной заменой этого документа» [48].

Сравнивая определения зарубежных и отечественных авторов, можно прийти к выводу, что в целом определения понятия «реферирование» совпадают. А.В. Бастриков и Д.Б. Клевеленд в своих определениях выделяют главную суть реферирования – передачу содержания первичного текста.

Разобравшись в понятии реферирования, следует перейти к следующему понятию, которое тесно связано с предыдущим, а именно, к понятию «реферат». Самое понятие «реферат» получило следующее определение: «краткое точное изложение содержания документа, включающее основные фактические сведения и выводы, без дополнительной интерпретации или критических замечаний автора реферата» [8]. Данное определение из ГОСТа во многом совпадает с определением, которое дал В.И. Соловьев, один из авторов популярной методики реферирования. Соловьев дал понятию «реферат» следующее определение: «Семантически адекватное, ограниченное малым объемом и вместе с тем возможно полное изложение основного содержания первичного документа» [9]. Данное определение понятия «реферат» относится к информационной литературе. Следующее определение «реферата» дано Л.П. Маркушевской и Ю.А. Цапаевой относительно реферирования научных текстов. В своем методическом пособии авторы дают понятию реферата следующее определение: «Реферат (от лат. «refero», что означает «сообщаю») представляет собой краткое изложение в письменном виде или в форме публичного доклада содержания научного труда (трудов), литературы по теме с раскрытием его основного содержания по всем затронутым вопросам, сопровождаемое оценкой и выводами референта» [24]. По мнению авторов реферат должен дать читателям объективное представление о характере освещаемого текста, изложить наиболее существенные моменты его содержания.

Проанализировав определения понятия «реферат», описанные выше, можно выделить основные требования, которые выдвигаются к качеству реферата: объективность, полнота отражения содержания первоисточника, включение основных фактических сведений, семантическая адекватность первичного и вторичного текстов.

1.1.3 Методы реферирования

Несмотря на многовековую историю и широкую востребованность, реферирование развивалось в основном эмпирически. Осознание необходимости создания теоретических основ и методик процесса реферирования пришло лишь во второй половине прошлого века. Такое озарение было связано с появлением реферативных журналов, о которых упоминалось выше, а также благодаря стремительному росту важности вторичных

документов в информационном обслуживании. Теоретические основы и методики были необходимы для того, чтобы повысить качество рефератов. В связи с этим появились определенные стандарты и требования, которые регулировали качество рефератов как официальных информационных документов.

Повышенная востребованность и рост популярности рефератов вызвали необходимость в разработке научно обоснованной методики составления реферативных текстов. В научном пространстве стали появляться первые теоретические и методические работы по тематике реферирования.

В выпускаемых работах были предложены методы реферирования научных текстов, которые можно поделить на две группы. К первой группе относятся методы, которые основаны на анкетировании. Сюда входит фасетное и поаспектное реферирование. Вторая группа включает в себя методы, которые основываются на анализе смысловой структуры первоисточника.

В первой группе наибольшее распространение получил метод поаспектного реферирования, который был предложен В.И. Соловьевым. Данный метод заключается в семантическом анализе научного текста с помощью его разбиения на отдельные содержательные аспекты. К таким аспектам можно отнести проблему и тему исследования, область знания, цель работы и полученные результаты, методы и условия проведенного исследования. В.И. Соловьев считает, что данный перечень аспектов является грамотным ориентиром для качественного реферата, так как он обеспечивает не только семантическую адекватность первичного и вторичного текстов, но и эффективное восприятие информации со стороны потенциального читателя [9]. Такой подход способствует формализации процесса реферирования текстов, а также помогает привести все реферативные тексты к общему стандарту.

Второй метод реферирования из первой группы был предложен А.И. Жолковой, которая назвала данный метод фасетным реферированием [15]. Фасетное реферирование основывается на анализе содержания конкретной сферы научной деятельности, который заключается в наложении на реферируемый текст сетки фасетов, соответствующей числу категорий, отражающих специфику данной области в различных ее аспектах. Основным понятием данной методики являются элементарные логические сообщения, создающие структурированный реферат (табличного формата). А.И. Жолкова считает, что

табличная форма реферата – это один из способов индексирования документов и структуризации фактографической информации.

При выявлении групп методов мы говорили о том, что методы первой группы основаны на анкетировании. Вышеперечисленные методы и правда можно рассматривать как своего рода анкетирование, так как референту предлагается «анкета» с «вопросами», которые помогут ему найти в реферируемом тексте соответствующую информацию. Анкетные виды реферирования ограничены рамками заранее заданной схемы, отражающей структуру предметной области, а также потребности специалистов. Данная ограниченность является основным достоинством таких методов реферирования, так как заранее заданный список вопросов не позволит референту вынести в реферат нерелевантную информацию.

Однако помимо достоинств у данных методов есть и свои недостатки, которые вызывают у референтов определенные трудности. Д.И. Блюменау в своей работе отмечает, что одним из ключевых недостатков данных методов можно считать обращение к интуиции референта, который должен выделить «ключевую информацию» из первичного текста, основываясь на своих знаниях и опыте. К недостаткам автор также относит стремление к детальности самой методики, которое может усложнить процесс реферирования [4]. Н.М. Нестерова и Н.А. Герте в своей работе высказывают свою точку зрения по отношению к методам реферирования, которые основаны на анкетировании. Авторы пишут, что данные методы в большей степени ориентированы на создание профильных, специализированных рефератов, так как в них отражается именно та информация, которая удовлетворяет потребности специалистов конкретной области знания [26].

Ко второй группе методов реферирования относятся методы выделения содержания первоисточника на основе анализа реферируемого текста. К данной группе методов можно отнести логико-психологический анализ, который был предложен А.А. Вейзе [5]. В эту же группу можно отнести информационно-классификационное реферирование. В его основу лег критерий точности информирования. Толчком для его развития послужил метод поаспектного реферирования, а именно разделение информации по степени значимости и полезности для потребителей. Такой метод реферирования опирается на содержательный и логический анализ первоисточника, который позволяет оценивать значимость и порядок следования той или иной информации. Благодаря такому подходу данный метод реферирования стал широко применяться в информационных центрах.

Выбор метода в реферировании не играет ключевой роли, так как основная задача референта заключается в установлении семантической адекватности реферата и оригинала. Адекватным в плане содержания рефератом можно считать такой реферат, в котором в сокращенной форме, но без искажений, воспроизводится основное содержание текста оригинала. Это говорит о том, что вопрос семантической адекватности напрямую связан с проблемой основного содержания текста. Данную проблему можно решить с помощью анализа внутренних закономерностей процессов понимания и восприятия текста, которые являются ключевыми в любом виде смыслового преобразования, реферирование не исключение. Понимание первоисточника при реферировании должно быть полным, иначе процесс реферирования может иметь случайный характер.

1.1.4 Виды рефератов

Понимание текста первоисточника тесно связано с его компрессией, а компрессия в свою очередь зависит от цели реферирования, которая напрямую зависит от вида реферата. Прежде чем перейти к основной части данного раздела, необходимо разобраться, что подразумевается под полным пониманием текста первоисточника. Полное понимание предполагает оценку всего текста оригинала, а это требует установления связей между всеми смысловыми частями текста и соотнесения этих частей с целым, чтобы не было никаких смысловых разрывов или пробелов. Результаты экспериментальных данных показывают, что смысловое восприятие текста и его понимание есть переход от воспринимаемых языковых средств к единицам внутреннего языка. Этот переход сопровождается выделением из текста ключевых элементов, которые А. Н. Соколов называет «смысловыми вехами». Смысловыми вехами могут быть любые слова из текста первоисточника, однако именно они представляют во внутренней речи большие смысловые отрезки, которые объединяются в «целостное мыслительное образование», несущее смысл первичного документа [27]. Таким образом, можно подытожить, что понимание является достаточно сложным аналитико-синтетическим процессом, который включает в себя анализ (дробление) и синтез (объединение), что делает его в то же время процессом смыслового свертывания.

Важно отметить, что любой текст изначально предрасположен к различного рода преобразованиям и сжатию. Такая предрасположенность основывается на особом свойстве текста – компрессионности. Данное свойство помогает референту сжимать текст с различной степенью компрессии в зависимости от цели реферирования, которая в свою очередь может зависеть от вида реферата.

Вместе с общими информативными рефератами востребованностью пользуются и так называемые специализированные рефераты. Название такого вида рефератов говорит само за себя, так как такие рефераты ориентированы на специалистов какой-либо конкретной предметной области. В случае специализированного реферирования Н.М. Нестерова и Н.А. Герте предлагают описывать процесс реферирования в терминах теории скопоса. Теория скопоса – это концепция из области переводоведения. Данная теория заключается в том, что при переводе в первую очередь должна учитываться функция целевого текста. При реферировании специализированных текстов Н.М. Нестерова и Н.А. Герте опираются на теорию скопоса, так как в данном виде рефератов именно целевая аудитория определяет отбор и комбинирование элементов содержания [26]. При этом, разумеется, меняется и критерий адекватности реферата. Теперь текст реферата должен быть адекватен не по отношению к тексту первоисточника, а по отношению к той коммуникативной ситуации, в которой ему предстоит функционировать.

Л.П. Маркушевская и Ю.А. Цапаева в своей работе указывают на то, что различают два вида рефератов: информативный реферат (реферат-конспект) и индикативный реферат (реферат-резюме). С первым видом мы уже знакомились выше, рассматривая виды рефератов в работах других авторов. Второй метод, который именуется индикативным, частично напоминает специализированный реферат, о котором также упоминалось выше.

Итак, Л.П. Маркушевская и Ю.А. Цапаева характеризуют информативный реферат как реферат, содержащий в обобщенном виде все основные положения текста первоисточника. Другими словами, информативный реферат – это реферат, содержащий фактографическую информацию (сведения о научных фактах). Например, если рассматривать реферирование научных текстов, то основными положениями будут являться сведения о методике исследования, использовании оборудования и сфере применения. В индикативном реферате, в отличие от информативного, указываются только основные аспекты содержания первичного документа. В таком виде реферата характеризуется содержание, информационное наполнение первичного документа. В нем называются только такие смысловые аспекты, которые заслуживают реферирования, т.е. содержат новые полезные сведения. Именно этим ориентиром на предметность текста индикативный реферат напоминает специализированный, который в свою очередь также направлен на реферирование текстов конкретной предметной области [24].

Однако, независимо от вида реферата, установка на реферирование предполагает обязательное полное осмысление текста первоисточника, так как моделирование

предметной ситуации позволит сохранить целостность денотатной структуры при существенном сжатии речевой формы. Это означает, что исходный текст (Т1) и текст реферата (Т2) представляют собой семантически адекватные речевые произведения, характеризующиеся разными логико-композиционными структурами и созданные на основе языковых единиц с разными объемами содержания. Адекватность исходного текста и текста реферата определяется на содержательном уровне, что вновь доказывает необходимость высокого уровня понимания текста первоисточника, так как без должного понимания будет невозможно добиться семантической равноценности двух текстов.

1.1.5 Сущность и методы компрессии материала первоисточника

При всем своем многообразии рефераты обладают некоторыми общими чертами. В рефераты не включаются рассуждения или исторические экскурсы, материал предстает в виде описания фактов, изложенных в тексте первоисточника. Во всех рефератах должна присутствовать точность и краткость изложения, информация не должна искажаться или обрастать субъективными оценками. Краткость изложения во многом достигается при помощи использования терминологической лексики, а также применения таблиц, формул и иллюстраций [24].

Реферат не должен представлять собой механический пересказ текста первоисточника. Текст реферата должен включать в себя все то, что заслуживает особого внимания со стороны потенциального читателя, при этом повторений и общих фраз при передачи смысла первоисточника стоит избежать. Кроме того, в реферат, как правило, не входит прямая речь и диалоги. В текст реферата целесообразно включать ключевые выводы автора первоисточника, если таковые имеются в реферируемом тексте.

Как неоднократно упоминалось выше, тексты рефератов характеризуются предельной точностью, которая достигается за счет экономной структуры предложения и правильного использования терминов. Зачастую для достижения краткости текста реферата термины записываются в сокращенной форме, так как это позволяет сэкономить место без особого ущерба для смысла. Сокращения терминологии могут быть как общепринятыми (например, adj. – прил.), так и типичными для конкретного текста первоисточника.

Для языка реферата свойственно использование определенных грамматико-стилистических средств. К таким средствам в первую очередь следует отнести простые законченные предложения, которые упрощают восприятия реферата. В передаче изложения текста первоисточника могут быть использованы причастные обороты, которые

обеспечивают экономию объема. Неопределенно-личные предложения позволяет сосредоточить внимание читателя только на важной информации, например, слова «анализируют, применяют, рассматривают и др.» являются акцентными и притягивают внимание на себя.

Язык реферата характеризуется большим количеством перечислений, которые неизбежно возникают в результате сжатия логического содержания реферируемого текста. В качестве повышения информативной и справочной роли референтом могут использоваться различные иллюстрации и схемы.

Объем реферата колеблется в зависимости от первичного печатного документа и характера реферата и может составлять 10-15% от объема первоисточника [24].

Сокращение объема материала достигается путем компрессии текста. Компрессия в свою очередь заключается в способности выражать одну и ту же мысль разными словами и речевыми оборотами. Любую мысль можно разделить на более значимые (главные) и второстепенные элементы, есть также причины и следствия, логические посылки и логические выводы. Перечисленные элементы составляют содержание всей мысли, однако реферат не требует передачи каждого из этих элементов, так как он призван передавать лишь наиболее существенную информацию. Таким образом, если хотя бы один из этих элементов не содержит в себе ключевой информации для передачи основного смысла содержания, то он не находит своего выражения в реферате и просто опускается.

Мы уже неоднократно говорили о том, что реферат призван передать основную мысль текста первоисточника, и это подводит нас к выделению основных задач в процессе реферирования. В ходе реферирования всегда выполняются две задачи: выделение основного смысла текста и краткое формулирование этого основного смысла текста. Таким образом, процесс реферирования заключается в двух внутренних процессах. К первому относится процесс отсеивания несущественного смысла, а ко второму относится перефразирование основной мысли в краткую форму речевого произведения.

Л.П. Маркушевская и Ю.А. Цапаева в своей работе экспериментально установили, что для успешного выполнения двух задач реферирования необходимо придерживаться следующей последовательности действий:

1. Необходимо произвести беглый осмотр текста первоисточника с целью ознакомиться с общим смыслом текста. Во время первичного осмотра обращается особое внимание на заголовки, графики, рисунки и др.

2. При вторичном просмотре текста производится вдумчивое чтение и ознакомление с общим содержанием, чтобы выстроить целостное восприятие текста.

3. Определяется основная тема текста первоисточника.

4. На данном этапе необходимо произвести смысловой анализ текста для того, чтобы отметить абзацы, содержащие ключевую информацию, которая раскрывает или уточняет заглавие текста.

5. Далее определяется ключевая мысль каждого отмеченного абзаца. Выстраивается логический план текста. Пункты плана необходимо формулировать назывными предложениями, так как потом их будет легче преобразовать в предложения, содержащие главную мысль абзаца [24].

1.2 Проблемы и методы автоматического реферирования текстов

1.2.1 Зарождение и становление идеи автоматического реферирования текста

Впервые о переизбытке существующей информации и проблеме ее хранения стали задумываться еще в середине прошлого столетия. Автоматическое реферирование текстов зародилось в начале 50-х годов прошлого века, но дисциплиной стало лишь к концу десятилетия в 1958 году после исследований в области научного реферирования текстов, которое было проведено Х.П. Луном [63]. Х.П. Лун уже в то время был обеспокоен постоянно растущим объемом информации, поэтому он предпринял попытки реферирования научных и технических текстов в области химии.

Люди из разных сфер жизни открыто высказывались о проблемах накопления большого массива информации. Одно из таких высказываний можно найти в статье Ричарда Фейнмана, профессора физики Калифорнийского технологического института и лауреата Нобелевской премии, которая была опубликована в журнале «Time» от 12 декабря 1960 года. В своей статье Р. Фейнман предсказал, что скоро на планете Земля произойдет информационный взрыв. Автор статьи предложил приз тому, кто сможет сократить одну страницу книги до одной двадцати пяти тысячных от оригинального объема, чтобы полученный текст можно было прочитать под микроскопом. Годом позже в своей сатирической статье Хэл Дрэппер высмеял идею о попытках справиться с предсказанной Р.

Фейнманом проблемой информационного взрыва, сократив данные до микроскопических размеров [84].

После новаторского исследования Х.П. Луна были написаны и другие основополагающие работы в данной области, включая исследования Г.П. Эдмундсона и Р.И. Уиллиса в 1961 и 1969 [52], [53] и работы других исследователей. В начале 1993 после недолгого затишья исследования возобновились благодаря работам К. Спарк-Джонс [74]. Исследования британского ученого помогли обратить внимание на существующую проблему и вызвали интерес общества к автоматическому реферированию.

Х. М. Торрес-Морено в своей книге выделяет ряд областей исследования, которые оказали влияние на становление и развитие реферирования. Автор перечисляет следующие области исследования: извлечение информации, поиск информации, машинное обучение, компьютерная наука, дискурс анализ, генерация естественного языка и искусственный интеллект. В настоящее время автоматическое реферирование текстов продолжает активно изучаться в нескольких областях, включая обработку естественного языка и другие смежные области [80].

Информационный взрыв был и остается реальной проблемой, однако человечеству удалось отсрочить время ее наступления. Экспоненциальный рост мощности новых электронных процессоров помогает преодолевать барьер хранения информации, наложенный старыми бумажными архивами.

Во времена Р. Фейнмана люди и подумать не могли, что появится удобный и легкий способ хранить информацию в электронном виде. Однако сейчас человечество дошло до того момента, когда даже интернет-пространство переполнилось огромным массивом данных, который сложно фильтровать и воспринимать. Кроме того, в Интернете существует информация на нескольких языках, что только увеличивает вышеупомянутые трудности, связанные с анализом документов. Автоматическое реферирование текстов призвано помочь человеку эффективно обрабатывать постоянно растущий объем информации, с которым люди просто физически не могут справиться.

Фактически, Международная корпорация данных (IDC) прогнозирует, что общий объем цифровых данных, ежегодно циркулирующих по всему миру, вырастет с 4,4 зеттабайта в 2013 году до 180 зеттабайт в 2025 году. Прогнозы Международной корпорации данных подтверждают постоянное увеличение потока текстовой информации, в связи с этим, задача автоматического реферирования становится все более актуальной для обработки естественного языка (NLP).

1.2.2 Автоматическое реферирование текста

1.2.2.1 Переход от «ручного» к автоматическому реферированию

Развитие методов реферирования осуществлялось в большинстве своем благодаря совершенствованию «ручного» реферирования, на что указывают различные исследования в области аналитико-синтетической обработки исходного текста [26]. Однако в современных реалиях информационного взрыва одним «ручным» реферированием не обойтись, поэтому еще в конце прошлого века люди начали прибегать к автоматизированным технологиям.

В настоящее время формализация реферирования стала одной из важнейших задач компьютерной обработки текстов. Исследователи, которые занимаются вопросом автоматического реферирования текстов, предлагают несколько способов обработки информации при помощи компьютера. Каждый из этих способов характеризуется своим методом выделения информации: квазиреферирование и краткое изложение содержание текста оригинала [26].

Квазиреферирование базируется на вычленении наиболее значимых и информативных элементов текста, на основе которых формируются квазирефераты. В настоящее время различают следующие методы отбора информации:

1. Статистические методы, которые основываются на оценке уровня информативности элементов исходного текста по частоте их появления в этом самом тексте. Именно частота появления служит ключевым критерием информативности слов или целых фрагментов.

2. Позиционные методы, которые заключаются главным образом в том, что информативность слова или предложения напрямую зависит от его положения в тексте.

3. Индикаторные методы, которые позволяют выделить в исходном тексте маркеры важности, характеризующие содержательную значимость текста [26].

Второй способ автоматического реферирования заключается в кратком изложении содержания. Такой способ основывается на выделении наиболее существенной информации при помощи автоматизированного качественного контент-анализа. Контент-анализ проходит в три стадии. Первая стадия заключается в сведении исходного текста к заданному числу фрагментов. В основе второй стадии лежит процесс поиска связей между

вычлененными фрагментами, по-другому их называют единицами значений. Третья стадия основывается на формировании выводов и обобщений, которые в дальнейшем входят в основу текста реферата в виде связанных смысловых единиц.

Выполнение такого рода преобразований требует использование манипуляций логическими предположениями и выделение шаблонов в имеющейся базе данных. В результате рождается концептуальная структура текста, т.е. концептуальные «выжимки» из текста [26].

Автоматическое реферирование несомненно ускоряет процесс реферирования в сравнении с «ручными» методами, а также исключает субъективность при изложении содержания текста первоисточника. Однако методы автоматического реферирования на сегодняшний день не удовлетворяют всем требованиям реферирования, в связи с чем качество рефератов остается достаточно низким. Причина низкого качества автоматического реферирования кроется в том, что при вычленении текстовых блоков не учитываются связи и отношения между ними, поэтому в результате реферирования порождаются тексты, которые лишены внешней связности и внутренней целостности.

1.2.2.2 Определение понятия «автоматическое реферирование»

В разделах выше уже было рассмотрено понятие реферирования, которое осуществляется человеком. Далее будет рассмотрено ключевое определение данной исследовательской работы – понятие «автоматическое реферирование». Сперва будут рассмотрены определения, которые даются в отечественных источниках, а затем в зарубежных. Т.В. Батура и А.В. Бакиева в своей монографии дают следующее определение: «Автоматическое реферирование (Automatic Text Summarization) — извлечение наиболее важных сведений из одного или нескольких документов и составление их краткого описания» [3]. В Википедии дается более лаконичное, но не менее точное определение понятия: «Автоматическое реферирование, аннотирование или суммаризация — создание краткой версии (реферата, аннотации) текстового документа с помощью компьютерной программы» [6]. Далее будут рассмотрены определения понятия «автоматическое реферирование», которые даются зарубежными авторами. Так, Т. Сакаи и К. Спарк-Джонс в своей работе дают следующее определение: «Автоматическое реферирование — это редуктивное преобразование текста первоисточника в текст реферата путем извлечения или генерации» [71]. Для определения понятия автоматического реферирования Х. М. Торрес-Морено обращается к Оксфордскому словарю [69], который определяет автоматическое реферирование как процесс создания сокращенной версии текста с помощью

компьютерной программы. Продукт, полученный от данного процесса, по-прежнему содержит наиболее важные пункты исходного текста [80]. Ключевое отличие автоматического реферирования от «ручного» заключается в том, что автоматическое производится с помощью компьютерной программы, на что в своих определениях указывают сразу два источника как отечественный, так и зарубежный.

1.2.2.3 Преимущества автоматического реферирования

Х. М. Торрес-Морено в своей работе приводит несколько веских причин в пользу автоматического реферирования документов:

1) Алгоритмы автоматического реферирования менее предвзяты, чем человек, выполняющий реферирование.

Порой тексты сопровождаются небольшими сводками или краткими изложениями содержания. Обычно такие сводки выполняются самим автором или специальными людьми, чьи услуги обходятся недешево и не всегда гарантируют качество. Качество человеческой работы может быть спорным, так как каждый человек – отдельная личность, со своими знаниями и опытом, на который он опирается, выполняя свою работу. Если дать один и тот же текст разным людям, то на выходе получатся разные сводки, так как каждый человек выделит те ключевые моменты и идеи, которые сочтет нужным. Алгоритм будет работать согласно заданным параметрам, он не будет опираться на свое видение или опыт, как человек.

2) Использование автоматических или полуавтоматических систем реферирования позволяет коммерческим службам подготовки аналитических обзоров увеличивать количество текстов, которые они могут обрабатывать.

Автоматическое реферирование оптимизирует скорость работы и является менее ресурсно-затратным аналогом человеческого труда [80].

1.2.2.4 Этапы и подходы к автоматическому реферированию

Автоматическое реферирование текста – это сложный процесс, который должен быть разбит на некоторые модули. Одним из таких модулей является предварительная обработка текста. Предварительная обработка позволяет преобразовать последовательность битов, то есть простой текстовый документ, в объект с минимальными языковыми особенностями, такими как слова и предложения. Предварительная обработка имеет две цели: нормализация слова и сокращение лексики в тексте.

В упрощенном виде (обязательная) предварительная обработка часто включает в себя следующие этапы:

1. Разбиение текста на сегменты, предложения, абзацы и т.д.
2. Разбиение сегментов на слова, или токенизация.
3. Нормализация слов (лемматизация, стемминг и т. д.).
4. Фильтрация стоп-слов.

Необязательная предварительная обработка может включать в себя:

1. Аннотация с помощью грамматических тегов или тегирование частей речи.
2. Распознавание именованной сущности.
3. Извлечение терминов и ключевых слов.
4. Весовые условия (с бинарным весом, числом вхождений).

На сегодняшний день принято выделять два основных подхода к автоматическому реферированию текстов:

1. Экстрактивный метод реферирования текстов (Extractive summarization). Данный метод заключается в экстракции ключевых фраз и идей из исходного текста, которые в дальнейшем добавляются в реферат без каких-либо изменений в порядке их появления в тексте.
2. Абстрактивный метод реферирования текстов (Abstractive summarization). Данный метод используется для сокращения исходного текста путем его перефразирования. Абстрактивный метод позволяет использовать в тексте реферата абсолютно новые фразы и предложения, которые обобщают и передают основную мысль исходного текста.

Х. М. Торрес Морено считает, что основное различие между абстракцией и экстракцией заключается в разных философиях реферирования. Согласно философии «хорошего ученика» (good student), глубокое понимание документа – это необходимое условие для составления качественного реферата, который сохраняет содержание текста и намерения автора. С другой стороны, в методе «плохого ученика» (bad student) просто определяются предложения, которые кажутся наиболее важными на первый взгляд (из-за их положения и наличия слов), а затем такие предложения извлекаются из текста.

Помимо абстракции и экстракции существует метод сжатия предложений. Сжатие предложений – это особый метод, который стремится сократить длину предложений.

Данный метод приводит к степени сжатия предложений около 33%, что сопоставимо со степенью сжатия текста человеком [80].

Одним из самых известных алгоритмов для экстрактивного реферирования является TextRank [67]. Данный метод разработан на основе другого не менее известного алгоритма, разработчиком которого является Google. Данный алгоритм носит название PageRank и служит он для оценки и ранжирования web-страниц по степени важности. Работа TextRank строится схожим образом, однако данный алгоритм сортирует по важности не web-страницы, а наиболее важные слова, фразы и предложения в исходном тексте. Отобранные по важности фрагменты в дальнейшем будут являться частью реферата. У TextRank имеется ещё один аналог с немного иной спецификой работы. Алгоритм EmbedRank выполняет поиск самых важных фрагментов исходного текста на основе моделей дистрибутивной семантики [72]. Экстрактивные алгоритмы реферирования текстов имеют ряд преимуществ перед абстрактными, так как первые легче реализуются благодаря тому, что фрагменты извлекаются из текста в первоначальном виде и вставляются в реферат в порядке упоминания в исходном тексте.

1.2.2.5 Проблемы автоматического реферирования, проверка качества

Любое реферирование текстовых документов представляет собой процесс сжатия исходного текста в краткий реферат данного текста. Сжатие информации включает в себя потерю некоего процента информации от исходного текста. Процесс реферирования считается автоматическим, когда он выполняется алгоритмом или компьютерной системой. Но какую именно информацию следует включить в краткое изложение, которое предстанет перед пользователем? Логично, что сгенерированное краткое содержание текста должно содержать наиболее важную и репрезентативную информацию из исходного документа. Но как измерить эту самую репрезентативность и значимость информации? Это один из ключевых вопросов, на которые пытаются ответить алгоритмы автоматического реферирования текста. На данный момент для измерения репрезентативности информации была разработана специальная формула, которая указана ниже.

Соотношение между длиной сводки и длиной исходного документа рассчитывается по степени сжатия τ :

$$\tau = \frac{|\text{Summary}|}{|\text{Source}|}$$

где $||$ указывает длину документа в символах, словах или предложениях, τ выражается в процентах.

Итак, какое же оптимальное значение для степени сжатия? Американский национальный институт стандартов рекомендует, чтобы сводка не превышала 250 слов. Желательно, чтобы степень сжатия $\tau = 10\%$ от общего объема текста. Однако есть и другие стандарты. Некоторые из них допускают возможность 15-30% [80].

Еще одной проблемой автоматического реферирования считается проблема оценки качества реферата, производимого автоматическими системами. Оценивание качества реферата – это открытая проблема, в ответ на которую научное сообщество может предложить лишь частичные пути решения. Как правило, оценка реферата текстовых документов классифицируется на внутреннюю оценку (intrinsic evaluation) и внешнюю оценку (extrinsic evaluation) [74].

Потенциальные пользователи рефератов зачастую достаточно прагматичны, так как они считают, что ценность и качество реферата должны измеряться в соответствии с его полезностью при выполнении разного рода задач. Несколько исследователей пытались провести внешнюю оценку качества рефератов, основываясь на задачах, которые были связаны с реферированием текстов. Так, К. МакКеон провел эксперимент, в котором сравнил работу людей, которые анализировали информацию из разнородного набора документов. Четыре группы людей получили либо оригинальные документы, либо автоматически сгенерированные рефераты, либо рефераты, произведенные другими людьми. Эксперимент показал, что люди, которым давали рефераты документов, отвечали на вопросы лучше, чем люди, которые работали с оригинальными текстами. Более того, чем качественнее выполнен реферат, тем более удовлетворительными были ответы [64].

Б. Дорр провел эксперимент, целью которого было выяснить зависит ли скорость понимания и усвоения информации от длины текста. Исследование показало, что участники эксперимента на 65% быстрее делают выводы по прочитанному реферату, чем по оригинальному документу [51]. Однако внешние оценки имеют прежде всего практическое применение, поэтому их достаточно тяжело обобщить, так как они слишком сильно зависят от одной конкретной задачи.

Внутренняя оценка рефератов может производиться тремя способами: вручную, полуавтоматически и автоматически. Каждый из перечисленных методов имеет свои преимущества и недостатки. Первый метод наиболее ресурсно-затратный, так как требует большого количества человеческого времени, потому что каждый реферат должен быть прочитан, оценен и подтвержден определенным количеством людей (судей). Более того, ручная оценка носит субъективный характер, поскольку между судьями могут быть значительные различия. Для того, чтобы выровнять субъективность в процессе оценивания рефератов, были разработаны полуавтоматические методы оценки. Такого рода методы

вычисляют показатели сходства между конкретным рефератом, который подлежит проверки на качество, с рефератами, которые были выполнены другими людьми. Третий и последний подход - это автоматические внутренние методы: реферат оценивается по отношению к исходному документу, а не по отношению к рефератам, которые были выполнены другими людьми. Меры статистической дивергенции (Дженсена-Шеннона или Кульбака-Лейблера), применяемые к распределению вероятностей, являются приоритетными в данном подходе. Данный подход становится все более популярным и постепенно проникает в научное сообщество благодаря своей независимости от других рефератов в процессе оценивания качества.

1.2.2.6 Варианты автоматического реферирования в зависимости от типов текстов

Автоматическое реферирование текстов широко используется в разных сферах жизни и применяется к разным категориям текстов. Х. М. Торрес-Морено выделяет следующие варианты применения автоматического реферирования:

1. Реферирование новостных текстов.
2. Реферирование блогов, твитов.
3. Реферирования веб-страниц.
4. Реферирование электронной почты.
5. Реферирование отчетов для бизнесменов, политиков, исследователей.
6. Реферирование выдержек из биографий.
7. Автоматическое извлечение и генерация заголовков.
8. Предметно-ориентированное реферирование (области медицины, химии и др).
9. Реферирование мнений [80].

Так, основываясь на разделении Х. М. Торрес-Морено, можно сделать вывод о том, что реферирование напрямую зависит от типа реферируемого текста. Именно поэтому в рамках данного исследования мы обращаемся к конкретному типу текстов – новостным текстам.

1.3 Особенности новостных текстов как объектов реферирования

1.3.1 Значимость массмедиа

Данная исследовательская работа направлена на анализ результатов работы алгоритмов реферирования новостных текстов. В настоящей работе отдается предпочтение именно новостным текстам, так как реферирование новостей на данный момент актуально,

как никогда. Во введении уже говорилось о том, что многие новостные агентства и порталы заинтересованы в автоматическом реферировании новостей, что доказывает востребованность реферирования именно такого типа текстов. Кроме того, новости появляются ежедневно и в огромном количестве от разных новостных агентств, порталов и групп в социальных сетях, что ведет к необходимости фильтровать огромный поток каждодневно поступающей информации. Таким образом, можно сделать вывод о том, что новости активно «засоряют» интернет пространство, ввиду своего постоянно растущего количества. Для того, чтобы реферировать новостные тексты, стоит сперва разобраться в специфике медийного жанра и медиалингвистике.

Т.Г. Добросклонская в своей работе «Медиалингвистика: системный подход к изучению языка СМИ» говорит о том, что «основной объём речепользования приходится сегодня именно на сферу массовой коммуникации» [13], что доказывает нужду в реферировании конкретно новостных текстов.

Солганик Г.Я. в своей работе «К определению понятий "текст" и "медiateкст"» [32] также указывает на то, что в современной языковой ситуации наряду с функциональными стилями особая роль принадлежит массовой информации. Она выдвигается на первый план по следующим критериям: сила влияния, масштабность, престижность и широта распространения.

1.3.2 Язык массмедиа

Становление массмедиа одной из основных сфер использования речи привело к концентрации внимания ученых на проблемах функционирования языка в области массовой коммуникации. Во второй половине XX века началось активное изучение свойств медиаречи, так как различные аспекты использования языка СМИ привлекли внимание российских и зарубежных ученых. Вопросы языкового обеспечения массовых коммуникационных процессов рассматривались в работах Г.В. Степанова, Д.Н. Шмелева, В.Г. Костомарова, О.А. Лаптевой, Г.Я. Солганика, Ю.В. Рождественского, А.Н. Васильевой, С.И. Сметаниной, Т.Г. Добросклонской и др.

Язык средств массовой информации обладает определенной целостностью и единством, что обусловлено особенностями употребления речи в сфере массовой коммуникации, заданностью форм и устойчивой тематической структурой. Язык СМИ характеризуется достаточной степенью функционально-стилевой разнородности, что отражает универсальность тематики, открытость и подвижность текста массовой информации в смысле его взаимодействия со всеми прочими сферами речеупотребления. Конструктивное единство и функционально-стилистическое разнообразие составляют

важнейшую особенность языка массовой информации, которая позволяет определить его как особый вид функционально-стилевых единств.

Периодом становления концепции языка СМИ можно считать конец XX века, в то же время и сам термин уже прочно вошел в научно-лингвистический обиход. В настоящее время под языком СМИ понимается не только корпус текстов, который входит в сферу массовой коммуникации, но и устойчивая внутриязыковая система, которая характеризуется определенным набором лингвостилистических свойств и признаков. Одной из основных характеристик данной внутриязыковой системы является соотношение вербальных и аудиовизуальных компонентов, специфическое для каждого средства массовой информации (печать, радио, телевидение или Интернет). Все вышеперечисленное и послужило основой для становления и развития новой отрасли языкознания – медиалингвистики.

К концу XX века общий объем исследования языка СМИ достиг «критической отметки», к тому времени уже сложились все необходимые условия для оформления накопленных знаний и опыта в области изучения языка СМИ в самостоятельное научное направление. Новое научное направление получило название медиалингвистика, в рамках которого предлагался комплексный подход к изучению языка СМИ.

1.3.3. Предмет медиалингвистики

О предмете медиалингвистики можно догадаться, исходя из самого термина, который объединяет в себе две стези: «медиа» (массмедиа) и «лингвистику». Лингвистика изучает язык, а масс медиа – это сфера, в которой функционирует этот самый язык. Таким образом, в качестве предмета медиалингвистики выступает изучение функционирования языка в сфере массовой коммуникации. То есть медиалингвистика занимается исследованием определенной сферы речеупотребления – языка массмедиа.

Со второй половины XX века можно наблюдать как средства массовой информации росли в геометрической прогрессии. Стремительное увеличение традиционных СМИ (печать, телевидение, радио) сопровождается бурным развитием информационных технологий, которое привело к созданию всемирной сети Интернет. Впоследствии во Всемирной паутине формируется глобальное информационное пространство. Информационные и телекоммуникационные технологии проникли и стали воздействовать на все сферы общественной жизни.

Помимо воздействия на общество, СМИ оказывают значительное влияние и на функционирование языка. Г.Я. Солганик в своей работе писал: «Телевидение, радио, газеты, журналы оказывают сильнейшее влияние на людей, формируют языковые вкусы,

воздействуют на литературные нормы» [32]. Стоит вновь обратить внимание на то, что в настоящее время массовая коммуникация стала одной из самых интенсивных сфер речеупотребления. Общий объем информации, который ежечасно появляется в Интернет пространстве и традиционных СМИ, неуклонно растет, что стимулирует рост интереса к этой стремительно развивающейся сфере со стороны науки.

1.3.3.1 Структура и содержание понятия «медiateкст»

Термин «медiateкст» впервые появился в англоязычной научной литературе в 90-х годах прошлого столетия. Однако термин быстро распространился как в зарубежных, так и в отечественных научных кругах. Быстрое закрепление концепции медiateкста объяснялось повышенным интересом исследователей к изучению особенностей функционирования языка в СМИ.

Исследованием медиаречи занимались такие зарубежные ученые, как Т. В. Дейк, Н. Фейерклаф, А. Белл, М. Монтгомери, Р. Фаулер.

В отечественной науке значительный вклад в развитие теории медiateкста внесли следующие ученые: В.Г. Костомаров, С.И. Бернштейн, Д.Н. Шмелёв, А.Н. Васильева, Г.Я. Солганик, Ю.В. Рождественский, Б.В. Кривенко, И.П. Лысакова, С.И. Трескова.

Впервые концепция медiateкста была сформулирована в работах Т.Г. Добросклонской. Основу концепции составляет органичное сочетание единиц вербального и медийного ряда.

Так, Т.Г. Добросклонская в своей работе отмечает, что А. Белл, известный исследователь языка СМИ, в своей книге «Approaches to Media Discourse» [47] пишет следующее: «Определение медiateкста выходит за рамки традиционного взгляда на текст как на последовательность слов, напечатанных или написанных на бумаге. Понятие медiateкста гораздо шире: оно включает голосовые качества, музыку и звуковые эффекты, визуальные образы – иначе говоря, медiateксты фактически отражают технологии, используемые для их производства и распространения». Отечественные ученые не остались в стороне и предложили свои варианты определения понятия «медiateкст». Емкое и достаточно точное определение было предложено М.Ю. Казак. Автор дает понятию следующее определение: «Интегративный многоуровневый знак, объединяющий в единое коммуникативное целое разные семиотические коды (вербальные, невербальные, медийные) и демонстрирующий принципиальную открытость текста на содержательно-смысловом, композиционно-структурном и знаковом уровнях» [18]. Г.Я. Солганик дает другое определение понятию «медiateкст», однако не менее точное: «Медiateкст – разновидность текста, принадлежащая массовой информации, характеризующаяся особым

типом автора (принципиальное совпадение производителя речи и ее субъекта), специфической текстовой модальностью (открытая речь, многообразное проявление авторского я), рассчитанная на массовую аудиторию» [32].

В концепцию медиатекста также входят устойчивые параметры, которые помогают описать любой медиатекст с точки зрения его производства, канала распространения и лингво-форматных признаков. К упомянутым выше параметрам относятся следующие:

1. Способ производства текста (авторский – коллегиальный).
2. Форма создания (устная – письменная).
3. Форма воспроизведения (устная – письменная).
4. Канал распространения (средство массовой информации – носитель: печать, радио, телевидение, Интернет).
5. Функционально-жанровый тип текста (новости, комментарий, публицистика (features), реклама) [13]

Динамика речеупотребления в СМИ достаточно активна, вследствие чего происходит систематичное жанровое движение, которое лишает данную единицу необходимого признака устойчивости. Существует множество подходов к функционально-жанровому описанию текстов СМИ, однако наиболее гибко сочетает устойчивую структуру и постоянную подвижность текстового материала классификация, разработанная в рамках теории медиалингвистики. Данная классификация выделяет четыре основных типа медиатекста:

- новости;
- информационная аналитика и комментарий;
- текст-очерк (любые тематические материалы);
- реклама.

Универсальность данной классификации заключается в том, что она, во-первых, основана на функционально-стилистической дифференциации языка, которая была построена на базе учения академика В.В. Виноградова о стилях языка и речи. Кроме того, данная классификация позволяет отражать реальную комбинаторику функций воздействия и сообщения в различных типах медиатекстов. Так, новости – это такие тексты, которые в наиболее полной форме реализуют одну из главных функций языка – сообщение, и одну из главных функций массовой коммуникации – информативную. Медиа аналитика и комментарий выполняют функцию сообщения, а также усиливают воздействие за счет выражения мнения и оценки. Тематическим материалам свойственно усиление воздействия

в его художественно-эстетическом варианте. Реклама совмещает в себе функцию воздействия как функцию языка, которая реализуется благодаря широкому разнообразию лингво-стилистических средств выразительности, и функцию воздействия как функцию массовой коммуникации, которая реализуется с помощью особых медиа технологий. Во-вторых, данная классификация обладает широкими описательными возможностями, так как она позволяет описать большую часть текстов СМИ, как с точки зрения основных форматных признаков, так и в плане особенностей реализации в тексте языковых и медийных функций.

6. Тематическая доминанта или принадлежность к тому или иному устойчивому медиатопику.

Содержательная характеристика текста также является существенным параметром типологического описания текстов, она позволяет определить тематическую доминанту или медиатопику. Медиатопика – это одна из устойчивых тем, которая регулярно освещается в СМИ. Средства массовой информации организуют и упорядочивают быстро меняющуюся картину мира при помощи устойчивой системы медиатопиков. К таким темам относятся следующие: бизнес, спорт, политика и др. Следует подчеркнуть значение лингвокультурного фактора, так как в текстах СМИ происходит своего рода наложение языковой и информационной картин мира, что проявляется в наборе постоянных тематических составляющих, которые характерны для той или иной страны, той или иной культуры. Так, одной из основных тем, которая активно освещается в британских СМИ, является частная жизнь королевской семьи [13].

1.3.3.2 Методы изучения медиатекстов

При изучении текстов СМИ применяется широкий спектр методов текстовой обработки: от традиционных методов системного и контент-анализа до анализа дискурсивного, социолингвистического и культурологического. Такое разнообразие применяемых методов обусловлено междисциплинарностью исследований медиаречи.

Так, тексты СМИ в настоящее время изучаются с точки зрения разных дисциплин: лингвистики, социологии, психологии, культурологии, стилистики и пр. Применение различных методов анализа текстов позволяет получить цельное представление о реальных свойствах медиаречи, ее стилистике, о способах взаимодействия вербального и медийного уровней.

По мнению Т. Г. Добросклонской наиболее эффективными методами изучения медиаречи являются следующие:

1. Методы лингвистического анализа.
2. Метод контент-анализа.
3. Метод дискурсивного анализа.
4. Метод критической лингвистики.
5. Метод когнитивного анализа.
6. Метод лингвокультурологического анализа. [13].

1.3.3.3 Типы медиатекстов: новости

Для медиаречи структурообразующими являются новостные тексты, которые считаются базовыми текстами СМИ. Именно базовые тексты наиболее широко представляют главные черты и особенности той или иной сферы речеупотребления. Новостные тексты считаются базовыми текстами СМИ на основе следующих факторов:

1. Новостные тексты наиболее полно реализуют одну из главных функций массовой коммуникации – информативную.
2. Новостные тексты имеют определенный формат, который выделяет их среди прочих типов текстов.
3. Новости на семантическом уровне характеризуются устойчивой макро- и микроструктурой, которая отражает тематическую организацию информационного потока.
4. На синтагматическом уровне новостные тексты характеризуются наличием большого количества клише и тематических словосочетаний, регулярным использованием текстообразующих элементов (ссылки на источники информации, фразы-связки и др.).
5. Новостным текстам характерна высокая повторяемость и частая воспроизводимость, что делает этот тип текстов особенно значимым. [13].

Совокупность всех вышеперечисленных факторов и позволяет рассматривать новостные тексты в качестве базовых текстов массовой информации.

1.3.3.4 Лингвомедийные свойства новостных текстов. Общая характеристика новостных текстов. Понятие новостной ценности

При описании содержательной стороны новостных текстов специалисты в области масс-медиа часто используют такое понятие, как «новостная ценность». Каждый день в мире случается огромное количество событий, но только некоторые из них освещаются средствами массовой информации. Освещаемые СМИ события отбираются по критерию

«новостной ценности». Новостная ценность события определяется целым рядом факторов: новизна, актуальность, пространственная или психологическая близость к получателю информации, возможные последствия для аудитории, фактор человеческого интереса и конфликтность. Мартин Монтгомери в своих работах выделяет следующие критерии новостной ценности:

1. Актуальность и новизна. Новости всегда должны быть свежими, в них должна быть представлена актуальная информация текущего дня.
2. Масштаб и значимость. Участники новостных событий со значимым социальным положением привлекают больше внимания со стороны зрителей или читателей. В новостях освещаются события связанные с жизнью известных людей: политиков, знаменитостей, спортсменов и др. Более того, особое значение имеет и масштаб события. Так, крупная катастрофа с большей вероятностью попадет в новости нежели небольшая авария.
3. Последствия события для массовой аудитории. По естественным причинам люди больше всего интересуются тем, что непосредственно влияет на их жизнь: повышение налогов, рост цен, погодные условия и др.
4. Пространственная близость. Сообщения, которые освещают события, происходящие в относительной близости от медиа аудитории – в стране, регионе, городе и т.д., имеют большую новостную ценность, чем те события, что происходят за пределами интересующей потребителя территории.
5. Конфликтность или негативный характер информации. В мировой журналистской практике широко распространено мнение о том, что «плохие новости» сильнее привлекают внимание людей, чем «хорошие». Человеческая психика устроена таким образом, что негативные сообщения о катастрофах, катаклизмах, конфликтах вызывают больший интерес со стороны медиа аудитории, чем «хорошие новости». Так, сообщения конфликтно-негативного характера отмечаются особой новостной ценностью.
6. Эмоциональный фактор. Среди новостных программ широко распространены сообщения, которые рассчитаны на эмоциональное воздействие. Для того, чтобы аудитория не «уставала» от сухих фактов, которые обычно освещаются в новостях, необходимо вызвать ее эмоции. Поэтому в новостных выпусках можно часто встретить такие сообщения, которые вызывают широкий спектр эмоций – сочувствие, сострадание, восхищение, умиление и др. Например, спасение животного из рук браконьеров, героический поступок человека и т.д. [68].

Британский лингвист Р. Фаулер в своей работе высказывается о новости следующим образом: «Новости являются репрезентацией мира в языке; поскольку язык представляет собой семиотический код, он навязывает определенную структуру ценностей социального и экономического происхождения, вне зависимости от того, что репрезентируется; и новости, таким образом, как любой дискурс, моделируют то, о чем сообщают. Новости являются репрезентацией именно в этом смысле; это не отражение фактов, лишенное ценностного компонента» [56]. Данное высказывание Р. Фаулера говорит о том, что новости, появляющиеся в СМИ, отражают картину нашего мира. Однако это не бездумное отражение протекающей вокруг нас реальности, это обдуманная и тщательно отобранная информация, которая навязывает определенные социальные и экономические ценности. Таким образом, можно сделать вывод, что новостная ценность коррелирует с ценностями, которые транслируются обществу через канал СМИ.

Классификация новостных сообщений по содержанию основывается на двух дихотомических категориях: «hard news – soft news» и «local news – foreign news». Также новостные тексты классифицируются по тематическим блокам: политика, экономика, происшествия, спорт, культура и др. Новости, относящиеся к категории «hard news», основаны на фактах и отвечают на вопросы что, где, когда, а новости из категории «soft news» основаны на человеческом интересе и ориентированы на то, чтобы вызвать эмоции.

Основу любого новостного текста составляет первая категория – «hard news», а сообщения из категории «soft news» служат для «разбавления» фактологической составляющей, чтобы вызвать самые разнообразные эмоции аудитории. Категория «local news» состоит из новостных сообщений о событиях в стране, регионе или городе, а сообщения из категории «foreign news» освещают зарубежные события. Содержание новостных сообщений соотносится также с устойчивыми медиатопиками, к которым относятся экономика, финансы и бизнес, политика, образование, наука, культура, здравоохранение, социальная проблематика, погода, спорт и др. Данные медиатопики в свою очередь могут подразделяться на более мелкие темы. Например, медиатопик «происшествия» делится на такие тематические единицы, как катастрофы, криминал и стихийные бедствия.

Цельный новостной текст состоит из сообщений разных содержательных категорий и тематических групп, которые разворачиваются в соответствии с требованиями новостного формата того или иного средства массовой информации.

1.3.3.5 Особенности формата новостных текстов в средствах массовой информации в сети Интернет

Термин «формат», активно используемый в масс медиа, означает устойчивое сочетание формальных признаков медиатекста с определенным содержанием.

Новостные тексты, которые распространяются через сеть Интернет, сочетают в себе все признаки «сетевого» текста с формально-содержательными особенностями печатных медиатекстов. Схожесть форматов объясняется тем, что Интернет тексты приближены к печатному первоисточнику и аналогично печатной прессе ориентированы на удобство зрительного восприятия. Именно поэтому большинство новостных текстов, представленных в сети Интернет, имеют форматные признаки, которые характерны для печатных новостных текстов. К таким признакам относятся следующие: презентация сообщения в краткой и развернутой формах, четкое распределение сообщений по содержательным категориям и тематическим группам, броские привлекающие внимание заголовки.

Как правило, новостная веб-страница выглядит следующим образом: в левой части сетевой страницы для удобства пользователя размещается устойчивый список медиатем, средняя полоса содержит краткое изложение наиболее интересных и актуальных на данный момент событий, справа размещены заголовки прочих сообщений о текущих событиях дня. Зачастую новостной текст в Интернете строится по принципу «перевернутой пирамиды» аналогично новостным текстам в газете. Краткие новостные сообщения представляют собой информационно-насыщенную часть пирамиды, которая в английской терминологии носит название «the lead», то есть первую фразу, содержащую всю ключевую информацию в сконцентрированном виде.

Особое значение при форматировании новостных текстов в Интернете имеет гипертекст. Высокоорганизованная система межтекстовых связей в сочетании с компьютерными технологиями создает огромные возможности для мгновенного информационного поиска, который может быть основан фактически на любом формально-содержательном компоненте текста, будь то отдельное слово или тематически связанное новостное сообщение [13].

Таким образом, можно сделать вывод, что при реферировании текстов стоит уделять особое внимание его жанровой принадлежности. Реферируя медиатексты сперва необходимо выбрать конкретный тип медиатекста для реферирования, так как у каждого типа свои особенности и структура. Нельзя, например, сравнивать реферирование новостных текстов с рекламными.

Кроме того, при реферировании новостных текстов стоит учитывать медиатопику. Не следует одновременно пытаться реферировать новостные тексты разной тематики, так как в них, как правило, будет использоваться разная тематическая лексика. Более того, необходимо учитывать тот факт, что медиатопики в свою очередь могут подразделяться на более мелкие темы. Например, новостные тексты из рубрики по происшествиям могут делиться на тексты о стихийных бедствиях, криминале и катастрофах.

Выводы по первой главе

На основе вышеизложенного можно сделать вывод о том, что реферирование зародилось достаточно давно, однако в научном плане реферирование стало рассматриваться лишь в 18 веке М.В. Ломоносовым, который является первым российским теоретиком реферирования. Первые реферативные журналы публиковали рефераты исключительно научных работ. Кроме того, долгое время реферирование изучалось только на примере научных текстов. Несмотря на то, что в современном обществе активно обсуждается проблема избыточности информации и тема реферирования регулярно поднимается на различных научных конференциях, информации о реферировании информационной литературы, в том числе новостных текстов, практически нет в свободном доступе.

В первой главе были рассмотрены и сравнены понятия «реферат» и «реферирование», которые даются разными авторами. На основе изложенных выше определений целесообразно выделить следующие требования к качеству реферата: объективность, полнота отражения содержания первоисточника, включение основных фактических сведений, семантическая адекватность первичного и вторичного текстов. Опираясь на данные требования будет разработана система оценки качества автоматического реферирования во второй главе данной исследовательской работы.

Современная проблема информационного взрыва требует эффективных действий. Исходя из сведений, которые были изложены в данной главе, можно сделать вывод, что только «ручным» реферированием в борьбе с избыточностью информации не обойтись. Еще в конце прошлого столетия стали широко применяться автоматизированные технологии, однако тема автоматического реферирования изучена не полностью и существующие алгоритмы автоматического реферирования нуждаются в совершенствовании и доработке, о чем будет свидетельствовать следующая глава данной исследовательской работы.

ГЛАВА 2

Лингвистический анализ результатов автоматического реферирования новостных текстов

2.1 Алгоритмы автоматического реферирования

Настоящая исследовательская работа посвящена анализу результатов работы алгоритмов реферирования новостных текстов.

В рамках исследования для анализа были отобраны два алгоритма автоматического реферирования текстов. При поиске алгоритмов мы столкнулись с определенными трудностями. В сети Интернет представлено не так много рабочих алгоритмов для автоматического реферирования текстов. В ходе поиска работающих алгоритмов были просмотрены алгоритмы как представленные в виде кода, так и веб-приложения. Так, многие программные коды оказались нерабочими, например, коды с ресурса medium.com [65], [66] и towardsdatascience.com [81]. Попытки запустить коды с вышеперечисленных ресурсов не увенчались успехом, так как при запуске были получены ошибки различного характера, которые не удалось устранить с целью наладить работу кода. Код с ресурса analyticsvidhya.com [44] оказался рабочим, но при этом процесс запуска кода требует хорошее техническое оборудование и большое количество времени. Так, при запуске кода для текстов на английском языке обучение модели заняло более одной недели. Обучение модели для текстов на русском языке заняло бы еще больше времени, так как подгружаемый файл с векторным представлением слов (word embeddings) для русского языка значительно превышает размеры файла для английского языка. Таким образом, данный код оказался слишком ресурсозатратным, так как запуск кода требует большого объема оперативной памяти и длительный период ожидания. В сети Интернет были найдены следующие программы для автоматического реферирования в режиме онлайн, которые в настоящее время не являются действительными: Sumplify [75], SweSum [76], Greatsummary [58], Textteaser [78]. Помимо нерабочих веб-приложения были найдены и рабочие, однако многие из них не подошли для данного исследования по разным причинам. Так, Free Text Summary Generator [59] подходит только для реферирования текстов на английском языке. Online summarize tool [79] хоть и работает на русскоязычных текстах, однако реферат получается не единым текстом, а отдельными наиболее значимыми предложениями в произвольном порядке, то есть алгоритм данного реферата выбирает 4-5 наиболее важных предложения из текста и вычленяет их в произвольном порядке, нумеруя каждое предложение. Подобный формат реферата не подходит для полноценного анализа, так как реферат не представляет собой единый текст. Esummarizer [54] и Automatic text

summarizer [45] предлагают пользователю ввести в поле настроек желаемое количество предложений, которое должно войти в реферат. Так как все новостные тексты разного размера, то при создании выборки из рефератов пришлось бы вручную контролировать количество предложений, что достаточно ресурсозатратно и вовсе не автоматизирует процесс получения готовых рефератов, так как необходимо было бы вручную просчитывать оптимальное соотношение предложений в оригинальном тексте и в тексте реферата. SMMRY [73] на текст любой длины выдает предупреждение, что текст слишком короткий. Причина данного предупреждения остается неясной.

Несмотря на долгий и тщательный поиск рабочего алгоритма автоматического реферирования текстов, нам удалось найти подходящие алгоритмы для данного исследования. Так, для исследования были отобраны два алгоритма: программный код на языке программирования Python [57] и веб-приложение Text compactor [77] для автоматического реферирования текстов.

Изначально программный код на языке программирования Python был написан для текстов на английском языке. Большинство алгоритмов будь то программные коды или готовые веб-приложения ориентированы на англоязычные тексты. Проблема адаптации англоязычных программных кодов является одним из пунктов актуальности данной исследовательской работы. Для того, чтобы найденный программный код заработал на русскоязычных текстах необходимо было подгрузить стоп-слова для русского языка из библиотеки NLTK и изменить пороговое значение.

Поручая компьютеру создание реферата, его необходимо обучить делать то же, что делал бы обычный человек, создавая реферат вручную. Компьютер должен суметь проделать следующие шаги для создания текста реферата:

- найти ключевые слова в тексте, фразы и предложения;
- найти в тексте менее важные единицы;
- составить из текстовых единиц двух первых типов смысловые единицы реферата;
- составить текст реферата из смысловых единиц [28].

Практически во всех существующих алгоритмах автоматического реферирования ключевые предложения и ключевые словосочетания и слова исходного текста являются основными смысловыми единицами реферата. Ключевые предложения в своей последовательной совокупности образуют текст реферата. Извлечение ключевых предложения из первичного текста относится к методам экстракции, который применяется в двух анализируемых нами алгоритмах. Ключевые словосочетания и слова используются компьютером для построения табличных рефератов.

Программный код на языке программирования Python соответствует тем шагам, которые были перечислены в работе М.М. Османовой. Далее структура программного кода будет рассмотрена более детально.

Для реализации автоматического реферирования в программном коде из библиотеки NLTK подгружаются два пакета: `nltk.corpus` и `nltk.tokenize`. Из первого пакета подгружается датасет заранее определенных стоп-слов русского языка. Из второго пакета подгружается токенизатор. Он делит текст на серию токенов. Существует три основных токенизатора – токенизатор слов, предложений и регулярных выражений. В программном коде будет использоваться только токенизатор слов и предложений. Следующим шагом после подгрузки необходимых для работы пакетов идет удаление стоп-слов и сохранение их в отдельном массиве слов. Далее создается таблица частотности слов, которая представляет собой словарь Python, в котором будет записываться информация о том, сколько раз каждое слово появляется в тексте после удаления стоп-слов. Словарь может быть использован для каждого предложения, чтобы узнать, какие предложения имеют наиболее релевантное содержание в общем тексте. Следующий шаг – присвоение оценки каждому предложению в зависимости от слов, которые оно содержит и данных таблицы частотности слов. Далее используется метод `sent_tokenize()` для создания массива предложений. Затем необходимо создать словарь, чтобы записывать в него оценку каждого предложения, далее данные из словаря будут использованы, чтобы сгенерировать текст реферата. Далее для того, чтобы сравнить оценки предложений необходимо найти среднюю оценку каждого предложения. Полученное среднее значение станет пороговым значением. Затем для создания текста реферата применяется пороговое значение, предложения по порядку записываются в текст реферата. С программным кодом на языке программирования Python можно ознакомиться в Приложении А (см. Приложение А).

Text Compactor – это бесплатный онлайн инструмент для реализации автоматического реферирования. Данный рефератор был разработан Китом Эдиберном и основан на другом онлайн рефераторе – Open Text Summarizer. На сайте рефератора имеется три окошка, которые может использовать пользователь. В первое окошко подгружается текст для реферирования, во второе окошко встроена шкала процентов, то есть пользователю дается возможность выбрать желаемый объем реферата, а в третьем окошке появляется текст реферата.

После того, как пользователь вставляет текст в первое окошко, веб-приложение вычисляет частоту каждого слова в тексте. Затем для каждого предложения высчитывается оценка на основе частоты встречаемости слов, которые содержатся в данных предложениях. Самым важным предложением считается предложение с наибольшей

частотой встречаемости слов. Создатель алгоритма предупреждает, что алгоритм может работать некорректно на слишком маленьких текстах, а также данный рефератор не рекомендуется использовать на текстах художественной литературы.

2.2 Критерии оценки качества реферата

2.2.1 Метрики BLEU и ROUGE для оценки качества автоматического реферирования

При проведении автоматического реферирования, как правило, качество проделанного реферирования оценивается тоже автоматически. Данная исследовательская работа направлена на разработку системы оценки эффективности методов автоматического реферирования, которая будет применяться человеком вручную. Прежде чем переходить к обсуждению критериев «ручного» оценивания, которое проводится человеком, необходимо осветить тему автоматического оценивания, которое проводится машиной.

Основными метриками, которые используются для автоматической оценки качества рефератов, являются BLEU и ROUGE. Изначально метрика BLEU была разработана для анализа качества машинного перевода, однако сейчас она активно используется наравне с метрикой ROUGE, которая была специально разработана для анализа качества автоматического реферирования.

BLEU измеряет точность (precision): сколько слов в сгенерированных машиной рефератах появилось в рефератах, выполненных человеком. ROUGE измеряет полноту (recall): сколько слов в рефератах, сделанных человеком, появилось в рефератах, сгенерированных машиной.

Точность системы в пределах класса – это доля документов, действительно принадлежащих данному классу относительно всех документов которые система отнесла к этому классу. Полнота системы – это доля найденных классификатором документов, принадлежащих классу относительно всех документов этого класса в тестовой выборке.

2.2.2 Критерии оценки «ручного» реферирования

Исходя из определений понятия «реферат» в теоретической главе данной исследовательской работы были выделены основные требования, которые выдвигаются к качеству реферата: объективность, полнота отражения содержания первоисточника, включение основных фактических сведений, семантическая адекватность первичного и вторичного текстов. Однако стоит учитывать, что определения понятия «реферат», на основе которых были выделены требования, изложенные выше, относятся к ручному реферированию текстов не только информационной литературы, но и научной. Еще одним

критерием оценки качества реферата, о котором говорилось в предыдущей главе, является избегание повторов. Так, реферат должен включать в себя важную информацию для потенциального читателя, но при этом желательно избегать повторов одних и тех же фраз или одного и того же смысла, так как повторы не несут в себе никакой значимости для текста, а лишь увеличивают объем реферата, который в свою очередь должен быть кратким и лаконичным.

Критерии оценки качества, которые выдвигаются в отношении ручного реферирования представлены во многих научных работах. Так, Е.Г. Губанова в работе «Рекомендации по написанию рефератов» [10], которая относится к ручному реферированию научных текстов, пишет о том, что критерии оценки качества рефератов традиционно делят на две группы: общие и частные.

К общим критериям относятся следующие:

- соответствие реферата теме;
- глубина и полнота раскрытия темы;
- адекватность передачи первоисточника;
- логичность, связность;
- доказательность;
- структурная упорядоченность (наличие введения, основной части, заключения, их оптимальное соотношение);
- оформление (наличие плана, списка литературы, культура, цитирования, сноски и т.д.);
- языковая правильность.

Частные категории относятся к конкретным структурным частям реферата: введению, основной части, заключению.

1. Критерии оценки введения:

- наличие сформулированных целей и задач работы;
- наличие краткой характеристики первоисточников.

2. Критерии оценки основной части:

- структурирования материала по разделам, параграфам, абзацам;
- наличие заголовка к частям текста и их удачность;
- проблемность и разносторонность в изложении материала;
- выделение в тексте основных понятий и терминов, их толкование;
- наличие примеров, иллюстрирующих теоретические положения.

3. Критерии оценки заключения:

- наличие выводов по результатам анализа;
- выражение своего мнения по проблеме.

Так как в рамках настоящего исследования проводится анализ рефератов, которые выполняются автоматическим путем, то не все критерии, применимые к ручным рефератам, могут применяться к автоматическим рефератам. Кроме того, стоит учитывать то, что данные критерии прописаны для научных текстов, поэтому не все из перечисленных критериев могут быть применимы к новостным текстам. Например, критерии из категории частных критериев не применимы для оценки качества рефератов новостных текстов, так как структура новости отличается от структуры научной работы. Такие составляющие новости, как введение, основная часть и заключение сильно отличаются от таких же составляющих научного текста.

Однако некоторые критерии из категории общих критериев вполне могут быть применимы для оценки качества рефератов в рамках данной исследовательской работы. Для начала обсудим те критерии из первой группы, которые не подходят. Первый пункт «соответствие реферата теме» не может быть применим для оценки рефератов новостных текстов, так как в нашем случае реферат основывается на одном тексте, который обрабатывается алгоритмом, и он не является рефератом нескольких текстов, как это бывает в случае научных рефератов. Следующий пункт «глубина и полнота раскрытия темы» подходит лишь от части, так как в случае с новостными текстами при составлении рефератов нет заданной темы, которой нужно придерживаться и раскрывать, реферат выполняется, основываясь на одном исходном тексте. Однако данный пункт можно адаптировать под реферат новостного текста. Вместо полноты раскрытия темы при анализе рефератов новостных текстов необходимо оценивать полноту отражения содержания текста первоисточника. Следующий пункт, который не подходит для оценки качества рефератов новостных текстов, «доказательность», так как, как правило, данный пункт относится к научным текстам, в которых можно доказывать результаты, выводы и др. Следующий пункт «оформление» не подходит для оценки качества новостных текстов, потому что реферат новостного текста не предполагает наличие списка литературы, сносок и т.д. Следующий пункт, который не подходит для оценки рефератов новостных текстов, «языковая правильность». Автоматическое реферирование анализируемых алгоритмов осуществляется по методу экстракции, лексическая составляющая текста реферата и исходного текста ничем не отличается, поэтому данный пункт не может быть применим при оценке качества автоматического реферирования новостных текстов.

Далее будут рассмотрены критерии оценки качества рефератов, которые подходят для анализа результатов автоматического реферирования новостных текстов. Как было

упомянуто выше, при анализе качества реферата важно оценивать полноту отражения содержания первичного текста. Несмотря на то, что реферат является краткой репрезентацией основного текста, он должен содержать в себе все ключевые, важные сведения. Кроме того, важным критерием оценки качества является адекватность передачи первоисточника. Не только человек, выполняя ручное реферирование, может ошибиться и допустить смысловые ошибки, но и машина. В силу того, что автоматическое реферирование выполняется с помощью экстракции ключевых предложений из текста, алгоритм может выбрать предложения из разных абзацев, которые не будут иметь логической связи, что может привести к искажению смысла. Следующим критерием, который вытекает из предыдущего, является логичность и связность текста. Следующий критерий – это структурная упорядоченность. У новостного текста есть своя структура, которая должна сохраняться и в тексте реферата. Новость состоит из заголовка, подзаголовка, лида (вступления), основного текста и заключения [41]. Все эти структурные компоненты должны отражаться в реферате в том случае, если они имеются в исходном тексте.

Таким образом, в рамках данного исследования все критерии можно поделить на две группы: структурные и содержательные.

К структурным критериям относится структурная упорядоченность.

К содержательным критериям относится полнота отражения содержания. Полнота содержания определяется ответами на основные вопросы, на которые должен отвечать текст новости. Так, согласно «Догме» газеты «Ведомости» в новостях должны освещаться следующие вопросы: Кто? Что? Сколько? Где? Почему? Когда? [14]. На сайте Школы журналистики говорится, что новостью принято считать информацию, дающую аудитории ответы на следующие вопросы: Кто? Что? С кем? Когда? Почему? Каким образом? [41]. Таким образом, текст реферата должен содержать в себе ответы на основные вопросы, которые освещаются в тексте первоисточника. Необходимо учитывать, что не в каждой новости могут быть освещены все перечисленные выше вопросы, так как события, освещаемые в новостях индивидуальны и могут не иметь тех или иных сведений.

Следующим критерием из группы содержательных критериев является **точность передачи данных**, при определении которой учитывается наличие или отсутствие случаев искажения информации, степень полноты представления информации, наличие или отсутствие разрывов логики/связности текста и наличие или отсутствие повторов.

Вышеописанные критерии оценки качества рефератов можно наглядно продемонстрировать следующим образом:

Структурные:

1. Структурная упорядоченность

Содержательные:

1. Полнота отражения содержания

2. Точность передачи данных:

- а) Наличие или отсутствие случаев искажения информации
- б) Степень полноты представления информации
- в) Наличие или отсутствие разрывов логики/связности текста
- г) Наличие или отсутствие повторов

Опираясь на указанные выше критерии оценки качества рефератов будет производиться анализ полученных текстовых выборок.

2.2.3 Анализ автоматически сгенерированных рефератов

2.2.3.1 Пример анализа автоматически сгенерированных рефератов на основе разработанных критериев

В качестве материала для анализа в рамках исследования были отобраны текстовые данные информационного агентства «ТАСС» из рубрики по происшествиям. Всего для анализа было отобрано 100 новостных текстов. В общей сложности получилось три текстовых выборки по 100 текстов каждая. Первая выборка состоит из оригинальных новостных текстов, вторую выборку представляют собой рефераты, полученные на основе применения первого алгоритма, а третья выборка состоит из рефератов, которые получились с помощью второго алгоритма.

Каждая новость из выборки первичных текстов проходила через два алгоритма: программный код на языке программирования Python и онлайн рефератор Text Compactor. После полученные рефераты сравнивались. Ниже будет представлен анализ нескольких показательных примеров автоматического реферирования текстов.

При анализе каждой последующей новости сначала рассматривается реферат, полученный при помощи первого алгоритма, а затем реферат, полученный при помощи второго алгоритма. Далее будет произведено сравнение работы двух алгоритмов.

Анализ рефератов будет производиться согласно разработанной в ходе исследования системе критериев оценки качества. Первая новость и ее реферат представлены в Таблице 1 (см. Таблица 1).

Таблица 1 – Первый пример новостного текста и его реферата, сгенерированного алгоритмом Text Compactor

Текст первоисточника	Текст реферата
СМИ: торнадо обрушился на Ухань Несколько человек пострадали, передает агентство "Синьхуа" ШАНХАЙ, 14 мая. /ТАСС/. Несколько человек пострадали в результате обрушения торнадо на город Ухань в провинции Хубэй (КНР). Об этом сообщило в пятницу информационное агентство "<u>Синьхуа</u>". По данным агентства, порывы шквалистого ветра достигали 23,9 м/с. Сообщается, что смерч повредил строения, были повалены деревья. Местные власти уточняют число пострадавших. Торнадо также обрушился 14 мая на провинцию Цзянсу. В результате один человек погиб, 21 пострадал, передает "Синьхуа" со ссылкой на местные власти. По их информации, в населенном пункте Шэнцзэ нарушено электроснабжение, разрушено несколько зданий.	СМИ: торнадо обрушился на Ухань Несколько человек пострадали, передает агентство "Синьхуа" ШАНХАЙ, 14 мая. Несколько человек пострадали в результате обрушения торнадо на город Ухань в провинции Хубэй (КНР). В результате один человек погиб, 21 пострадал, передает "Синьхуа" со ссылкой на местные власти. По их информации, в населенном пункте Шэнцзэ нарушено электроснабжение, разрушено несколько зданий.

Приведенный выше реферат в целом соответствует структуре новости, сохраняя в себе все структурные элементы новости за исключением последнего элемента – заключения. Однако, как можно заметить, в тексте первоисточника тоже отсутствует смысловое заключение новости, так как создается ощущение незавершенного текста в виду отсутствия общих слов о происшествии в Китае или дальнейших действий относительно ситуации. Кроме того, стоит отметить, что при автоматическом реферировании машина не

разделяет текст на смысловые абзацы. Абзацы сохраняются относительно оригинального текста. Так, например, в тексте реферата после лида начинается основная часть новости, которая должна выноситься в отдельный абзац, но так как данное предложение в тексте первоисточника находится внутри абзаца, а не начинается с красной строки, то и в тексте реферата это предложение будет находиться внутри абзаца. Таким образом, нарушается лишь визуальная структура текста новости.

В целом в тексте реферата отражены основные содержательные компоненты новостного текста. Однако информация о том, сколько человек пострадало искажена, а информация о последствиях происшествия представлена не в полном виде. Так, в начале новости говорится про город Ухань в провинции Хубэй и дается информация о том, что число пострадавших уточняется властями. При этом в тексте реферата извлеклась информация о числе пострадавших и погибших, которая относится к провинции Цзянсу. Таким образом, текст реферата не в полной мере освещает необходимые вопросы, а также характеризуется отсутствием адекватной передачи данных первоисточника. В тексте реферата нарушена семантическая адекватность, так как происходит искажение смысла, который заложен в тексте первоисточника. Данная проблема возникла из-за того, что автоматическое реферирование производится с помощью экстракции предложений из оригинального текста. Алгоритм посчитал предложение с информацией о количестве пострадавших и погибших существенно важной, что является верным, однако данная информация относилась к другой провинции, которая не попала в текст реферата. Кроме того, как было замечено выше, информация о последствиях разрушения представлена не в полном виде. Так, информация об ущербе, который был нанесен городу Ухань, не попала в текст реферата, при этом информация о последствиях торнадо в населенном пункте Шэнцзэ, провинция Цзянсу, попала в текст реферата. Вследствие того, что часть информации вошла в текст реферата, а часть была утеряна, нарушается логика текста реферата, так как в лиде говорится о городе Ухань, а затем идет информация о разрушениях в другом населенном пункте.

В Таблице 2 (см. Таблица 2) представлена та же новость, но реферат выполнен вторым алгоритмом.

Таблица 2 – Первый пример новостного текста и его реферата, сгенерированного программным кодом на языке программирования Python

Текст первоисточника	Текст реферата
СМИ: торнадо обрушился на Ухань	СМИ: торнадо обрушился на Ухань

Несколько человек пострадали, передает агентство «Синьхуа» ШАНХАЙ, 14 мая. /ТАСС/. Несколько человек пострадали в результате обрушения торнадо на город Ухань в провинции Хубэй (КНР). Об этом сообщило в пятницу информационное агентство «Синьхуа». По данным агентства, порывы шквалистого ветра достигали 23,9 м/с. Сообщается, что смерч повредил строения, были повалены деревья. Местные власти уточняют число пострадавших. Торнадо также обрушился 14 мая на провинцию Цзянсу. В результате один человек погиб, 21 пострадал, передает «Синьхуа» со ссылкой на местные власти. По их информации, в населенном пункте Шэнцзэ нарушено электроснабжение, разрушено несколько зданий.	Несколько человек пострадали, передает агентство «Синьхуа» ШАНХАЙ, 14 мая. /ТАСС/. Об этом сообщило в пятницу информационное агентство «Синьхуа». По данным агентства, порывы шквалистого ветра достигали 23,9 м/с. Сообщается, что смерч повредил строения, были повалены деревья. Местные власти уточняют число пострадавших. Торнадо также обрушился 14 мая на провинцию Цзянсу. В результате один человек погиб, 21 пострадал, передает «Синьхуа» со ссылкой на местные власти. По их информации, в населенном пункте Шэнцзэ нарушено электроснабжение, разрушено несколько зданий.
--	---

В данном реферате нарушается структура новости, так как в текст реферата не попал лид новости, при этом остальные элементы новости сохраняются согласно тексту первоисточника.

Реферат освещает основные вопросы новости, не искажая данные, как это было в предыдущем реферате. Небольшое нарушение семантической адекватности можно наблюдать в первом предложении после подзаголовка. Данное предложение должно быть связано с лидом, но так как лид не вошел в текст реферата это предложение теряет свою смысловую связность в тексте. Кроме того, из-за опущенного лида данное предложение

является неуместным повтором заголовка. Опущенный лид также может исказить восприятие текста реферата. Так, если рассматривать текст реферата в отрыве от заголовка и подзаголовка, то потенциальному читателю не будет понятно, где именно был шквалистый ветер. Однако, если смотреть на заголовок, то станет понятно, где случилось стихийное бедствие.

В данном реферате в полной мере отражено содержание первоисточника и включены все основные фактические сведения. В отличие от предыдущего реферата в данный реферат попала информация о двух местностях, где было торнадо, а также попала информация о разрушениях как в Ухане, так и в Шэнцзэ.

Несмотря на то, что реферат второго алгоритма не исказил данные и хорошо передал содержание первоисточника, стоит учитывать тот факт, что реферат получился достаточно большим, так как в него вошли почти все предложение оригинального новостного текста.

В Таблице 3 (см. Таблица 3) представлен следующий новостной текст и его реферат.

Таблица 3 – Второй пример новостного текста и его реферата, сгенерированного алгоритмом Text Compactor

Текст первоисточника	Текст реферата
Женщина, пытавшаяся похитить деньги у родственника Сталина, пойдет под суд в Москве Подозреваемая и ее сообщники пытались похитить 3,4 млн рублей МОСКВА, 14 мая. /ТАСС/. Дело о попытке кражи 3,4 млн рублей у племянника Иосифа Сталина и ветерана Великой Отечественной войны Леонида Аллилуева направлено в суд. Об этом сообщили ТАСС в пресс-службе прокуратуры Москвы. "Таганская межрайонная прокуратура Москвы утвердила обвинительное заключение по уголовному	Женщина, пытавшаяся похитить деньги у родственника Сталина, пойдет под суд в Москве Подозреваемая и ее сообщники пытались похитить 3,4 млн рублей МОСКВА, 14 мая. 159 УК РФ ("Покушение на мошенничество в особо крупном размере)", - говорится в сообщении. В марте женщина и ее сообщники, зная, что на счетах 92-летнего Аллилуева есть крупная сумма, несколько раз звонили ему, представляясь правоохранителями, и говорили о необходимости провести судебную экспертизу по имеющимся у него 3,4 млн рублей. Обманутый ветеран снял со

делу в отношении 31-летней жительницы столицы. Она обвиняется в совершении преступления, предусмотренного ч. 3 ст. 30 ч. 4 ст. 159 УК РФ ("Покушение на мошенничество в особо крупном размере")", - говорится в сообщении. В марте женщина и ее сообщники, зная, что на счетах 92-летнего Аллилуева есть крупная сумма, несколько раз звонили ему, представляясь правоохранителями, и говорили о необходимости провести судебную экспертизу по имеющимся у него 3,4 млн рублей. Обманутый ветеран снял со счета деньги, чтобы передать их мошеннице, но вовремя понял, что его обманывают, и обратился в полицию. Женщину задержали в квартире потерпевшего. Дело в отношении нее выделено в отдельное производство, она содержится под стражей. Дело рассмотрит Таганский районный суд Москвы.	счета деньги, чтобы передать их мошеннице, но вовремя понял, что его обманывают, и обратился в полицию. Женщину задержали в квартире потерпевшего.
--	---

В полученном реферате нарушена структура новости, так как присутствуют все структурные элементы за исключением – лида. На месте лида находится обрывочное предложение, которое начинается с номера УК РФ. Данное предложение обрывочно извлеклось по причине того, что перед номером УК РФ идут сокращенные слова, алгоритм принял эти слова за отдельные самостоятельные предложения и не счел их значимыми для того, чтобы включить в текст реферата. Таким образом, предложение извлеклось не полностью, вследствие чего нарушается логика и связность текста. В тексте реферата отражены основные содержательные компоненты новости, которые соответствуют тексту первоисточника. Искажения информации, наличие повторов или неполнота представления данных не наблюдается.

В Таблице 4 (см. Таблица 4) представлен реферат, полученный на основе применения второго алгоритма, который представлен программным кодом на языке программирования Python.

Таблица 4 – Второй пример новостного текста и его реферата, сгенерированного программным кодом на языке программирования Python

Текст первоисточника	Текст реферата
Женщина, пытавшаяся похитить деньги у родственника Сталина, пойдет под суд в Москве Подозреваемая и ее сообщники пытались похитить 3,4 млн рублей МОСКВА, 14 мая. /ТАСС/. Дело о попытке кражи 3,4 млн рублей у племянника Иосифа Сталина и ветерана Великой Отечественной войны Леонида Аллилуева направлено в суд. Об этом сообщили ТАСС в пресс-службе прокуратуры Москвы. "Таганская межрайонная прокуратура Москвы утвердила обвинительное заключение по уголовному делу в отношении 31-летней жительницы столицы. Она обвиняется в совершении преступления, предусмотренного ч. 3 ст. 30 ч. 4 ст. 159 УК РФ ("Покушение на мошенничество в особо крупном размере")", - говорится в сообщении. В марте женщина и ее сообщники, зная, что на счетах 92-летнего Аллилуева есть крупная сумма, несколько раз звонили ему, представляясь правоохранителями, и	Женщина, пытавшаяся похитить деньги у родственника Сталина, пойдет под суд в Москве Подозреваемая и ее сообщники пытались похитить 3,4 млн рублей МОСКВА, 14 мая. /ТАСС/. Дело о попытке кражи 3,4 млн рублей у племянника Иосифа Сталина и ветерана Великой Отечественной войны Леонида Аллилуева направлено в суд. Об этом сообщили ТАСС в пресс-службе прокуратуры Москвы. Она обвиняется в совершении преступления, предусмотренного ч. 3 ст. 30 ч. 4 ст. 159 УК РФ ("Покушение на мошенничество в особо крупном размере")", - говорится в сообщении. Обманутый ветеран снял со счета деньги, чтобы передать их мошеннице, но вовремя понял, что его обманывают, и обратился в полицию. Женщину задержали в квартире потерпевшего. Дело в отношении нее выделено в отдельное производство, она содержится под стражей. Дело рассмотрит Таганский районный суд Москвы.

говорили о необходимости провести судебную экспертизу по имеющимся у него 3,4 млн рублей. Обманутый ветеран снял со счета деньги, чтобы передать их мошеннице, но вовремя понял, что его обманывают, и обратился в полицию. Женщину задержали в квартире потерпевшего. Дело в отношении нее выделено в отдельное производство, она содержится под стражей. Дело рассмотрит Таганский районный суд Москвы.	
---	--

Полученный реферат в полной мере отражает структуру исходного текста, все структурные компоненты новости находятся на своем месте. Несмотря на то, что структура новости сохранена, в основной части текста наблюдается нарушение логики, так как в предложении *«Она обвиняется в совершении преступления, предусмотренного ч. 3 ст. 30 ч. 4 ст. 159 УК РФ ("Покушение на мошенничество в особо крупном размере")»*, - говорится в сообщении» непонятно, кто именно обвиняется в преступлении, потому что предложение, в котором дается имя преступнику не попало в текст реферата.

В структурном плане второй алгоритм проявил себя лучше первого, однако в рефератах обоих алгоритмов присутствовало нарушение логики.

2.2.3.3 Результаты анализа автоматически сгенерированных рефератов на основе разработанных критериев

Выше были рассмотрены два новостных текста из текстовой выборки, которая собиралась в рамках исследования для анализа. В качестве примера был продемонстрирован процесс анализа, который проводился с каждым новостным текстом из выборки. Приведенные выше рефераты и результаты их анализов наглядно показывают, что на данный момент времени автоматическое реферирование не находится на должном уровне, чтобы производить качественные рефераты.

В ходе проведения анализа всей выборки была составлена таблица с критериями оценки качества и подсчитанным количеством ошибок в каждой из категории критериев

относительно двух алгоритмов. С полными результатами анализа можно ознакомиться в Таблице 5 (см. Таблица 5).

Таблица 5 – Количество нарушений норм реферирования, признаков, определяющих качество реферата

Критерии		Первый алгоритм	Второй алгоритм
		Кол-во	Кол-во
Структурные:	Структурная упорядоченность	48	24
Содержательны:	Полнота отражения содержания	11	4
	Точность передачи данных:	1	3
	а) Наличие или отсутствие случаев искажения информации	21	13
	б) Степень полноты представления информации	40	37
	в) Наличие или отсутствие разрывов логики/связности текста	6	14
	г) Наличие или отсутствие повторов		

Для начала рассмотрим результаты анализа по первой выборке, тексты которой были получены через онлайн рефератор Text Compactor. Больше всего ошибок было насчитано по критерию структурной упорядоченности, их количество равняется 48. Данный показатель говорит о том, что практически в половине текстов выборки были обнаружены проблемы в структуре новостного текста реферата. В основном проблемы в структуре

наблюдались в лиде. Проблема подобного рода, по нашим предположениям, вызвана тем, что лид зачастую повторяет собой заголовок и подзаголовок, поэтому алгоритм во избежание повторов не включал лид в текст реферата.

Второе место по количеству ошибок занимает критерий наличия и отсутствия разрывов логики/связности текста. По данному критерию было обнаружено 40 ошибок. Если говорить о значимости данных ошибок, то критерий структурной упорядоченности уступает в значимости критерию наличия разрывов логики/связности текста, так как наличие структуры новостного текста в реферате не влияет на смысловую передачу данных в то время, как нарушение логики напрямую влияет на смысл, который заложен в тексте реферата. Далее будет приведен пример разрыва логики/связности текста: *«"Дознавателем МЧС было возбуждено уголовное дело по ст. 168 УК РФ (Уничтожение или повреждение имущества по неосторожности). Прокуратурой региона материалы изъяты и направлены в полицию для дальнейшего расследования", - сообщили в пресс-службе»* → *«ГУ МЧС по Краснодарскому краю в пятницу сообщало, что в Кореновске загорелся строящийся детский сад, площадь пожара составила 300 кв. Прокуратурой региона материалы изъяты и направлены в полицию для дальнейшего расследования", - сообщили в пресс-службе»*. Сначала представлен отрывок из оригинального новостного текста, а затем располагается отрывок из реферата. Как можно заметить, в текст реферата не вошло первое предложение, указанное в отрывке оригинального текста, вследствие чего возник разрыв логики, так как в тексте реферата непонятно, какие именно материалы были изъяты прокуратурой региона. Рассмотрим еще один пример: *«ТАСС, 14 мая. Суд в Саранске заключил под стражу бывшего заместителя председателя правительства Мордовии - министра целевых программ Алексея Меркушкина, обвиняемого в даче взятки в размере 7 млн рублей. Об этом сообщили в пятницу ТАСС в пресс-службе Следственного управления СК РФ по Мордовии»* → *«ТАСС, 14 мая. Об этом сообщили в пятницу ТАСС в пресс-службе Следственного управления СК РФ по Мордовии»*. В данном примере можно наблюдать разрыв логики и связности текста, который возник в результате отсутствия лида. В текст реферата не вошло первое предложение отрывка оригинального текста, которое является лидом новости. Таким образом, предложение, которое вошло в текст реферата выглядит неуместно и вызывает вопросы, так как из текста реферата непонятно, о чем именно сообщили ТАСС в пресс-службе.

Существенное количество ошибок было также обнаружено по критерию степени полноты представления информации. Всего была насчитана 21 ошибка. Далее будет рассмотрен пример неполноты представления информации: *«По данным надзорного*

ведомства, трое несовершеннолетних в пятницу решили прокатиться по Зеленограду на автомобиле. В результате водитель машины, 14-летний юноша, выезжая из двора, не справился с управлением и врезался в дерево. Один из подростков госпитализирован» → «По данным надзорного ведомства, трое несовершеннолетних в пятницу решили прокатиться по Зеленограду на автомобиле». Данный пример показывает то, как в тексте реферата опускается существенно важная информация, представленная в оригинальном тексте новости. Предложение о том, что трое несовершеннолетних решили прокатиться на автомобиле, не говорит о том, что произошло дальше, каким был итог их поездки.

Следующим критерием по количеству ошибок является критерий полноты отражения содержания. По данному критерию было насчитано 11 ошибок. Такой результат говорит о том, что практически 90% текстов из выборки соответствовали содержанию первичного текста и отвечали на основные вопросы, на которые опирается новость.

Следующим критерием по количеству ошибок является критерий наличия или отсутствия повторов в тексте. По данному критерию было насчитано всего 6 ошибок. Далее будет рассмотрен пример по данному критерию: *«Объем разлившихся нефтепродуктов в районе реки Колва на Ошском месторождении в Ненецком автономном округе (НАО) составил не менее 7 т, говорится в пятницу в сообщении пресс-службы прокуратуры региона. "По предварительным данным, объем вытекания нефтесодержащей жидкости составил не менее 7 т", - говорится в сообщении».* В данном случае представлен отрывок из текста реферата, в котором наглядно видно, что информация о том, что разлилось не менее 7 тонн нефтепродуктов попала в реферат дважды. Второй раз данная информация уже не является релевантной и лишь занимает место в тексте реферата, который должен быть кратким, лаконичным и содержать основную суть первоисточника.

Наименьшее количество ошибок было обнаружено по критерию наличия или отсутствия случаев искажения информации. Всего была найдена 1 ошибка, пример которой был приведен в первом новостном тексте, анализ которого был представлен выше. Ошибка заключалась в искажении информации о количестве пострадавших.

Далее будут рассмотрены результаты анализа по второй выборке, тексты которой были получены с помощью использования программного кода на языке программирования Python. Наибольшее количество ошибок было найдено по критерию наличия или отсутствия разрывов логики/связности текста. По данному критерию было насчитано 37 ошибок. Рассмотрим пример наличия разрыва логики или связности текста: *«ГУ МЧС по Краснодарскому краю в пятницу сообщало, что в Кореновске загорелся строящийся*

детский сад, площадь пожара составила 300 кв. м.» → «м.». В данном примере видно, что алгоритм принял сокращенное слово «метры» за самостоятельное предложение и вынес его в текст реферата в отрыве от остальной части предложения. Таким образом, нарушилась логика и связность текста реферата.

Следующим критерием по количеству ошибок является критерий структурной упорядоченности. По данному критерию было обнаружено 24 ошибки. Аналогично первому алгоритму, второй алгоритм нарушал структуру новостного текста, исключая из текста реферата лид.

Практически одинаковое количество ошибок было найдено по критерию степени полноты представления информации и критерию наличия или отсутствия повторов. Всего было насчитано 13 и 14 ошибок соответственно. Сначала будет рассмотрен пример неполноты представления информации: *«"Из Северо-Уральского межрегионального управления Росприроднадзора поступила информация об обнаружении разлива нефтепродуктов площадью 15 000 кв. м. на территории земельного участка Пуровского района Ямало-Ненецкого автономного округа (земли лесного фонда Ноябрьского лесничества) - Карамовское месторождение АО "Газпромнефть-Ноябрьскнефтегаз", - говорится в сообщении»* → *«"Из Северо-Уральского межрегионального управления Росприроднадзора поступила информация об обнаружении разлива нефтепродуктов площадью 15 000 кв.»*. На примере данных отрывков можно заметить, что алгоритм посчитал точку после сокращенного обозначения слова «квадратных» окончанием предложения, а информацию, которая находилась в оставшейся части предложения неважной, так как она не была включена в текст реферата. Таким образом, логика изложения информации была сохранена, однако сама информация представлена не в полном виде, так как предложении оборвалось. Рассмотрим еще один наглядный пример неполноты представления информации: *«Ранее сообщалось, что подсудимый, занимая должность мэра, получил от местного предпринимателя для заключения договора подряда и общего покровительства денежные средства в размере 500 тыс. рублей, являющихся частью требуемой взятки в общей сумме в 7 млн рублей»* → *«Ранее сообщалось, что подсудимый, занимая должность мэра, получил от местного предпринимателя для заключения договора подряда и общего покровительства денежные средства в размере 500 тыс.»*. В данном примере происходит аналогичная ситуация, что и в примере выше. Алгоритм неверно определяет границы предложения, вследствие чего важная смысловая часть предложения не попадает в текст реферата.

Далее будет рассмотрен пример наличия повторов в тексте реферата: *«Анонимное сообщение с угрозой взрыва поступило в гимназию №27 Казани. "Поступило анонимное сообщение о взрывном устройстве в гимназии №27", - рассказал собеседник ведомства. В мэрии города в свою очередь сообщили, что угроза взрыва оказалась ложной. Правоохранительные органы провели проверку в гимназии №27, взрывных устройств не обнаружено. "Как показало обследование, вызов оказался ложным", - говорится в сообщении»*. В данном случае представлен отрывок из реферата. Из примера видно, что отрывок перегружен повторами: два раза повторяется информации о том, что поступило анонимное сообщение об угрозе взрыва, а также дублируется информация о том, что вызов оказался ложным.

Следующим критерием по количеству ошибок является критерий полноты отражения содержания. По данному критерию было насчитано 4 ошибки. Такой результат говорит о том, что 96% текстов из выборки соответствовали содержанию первичного текста и отвечали на основные вопросы, на которые опирается новость.

Наименьшее количество ошибок было насчитано в критерии наличия или отсутствия искажений. Всего было насчитано 3 ошибки. Однако сам критерий является достаточно серьезным и показательным. Рассмотрим примеры искажения информации: *«Один из осужденных полицейских - Николай Плебух - свою вину в суде не признал, считая, что действовал правомерно. Вечером 19 октября 2019 года в отдел полиции по Московскому району Калининграда за совершение административного правонарушения был доставлен 36-летний мужчина. Сообщалось, что он, находясь в нетрезвом состоянии, повредил радиатор батареи в камере для задержанных и получил сильные ожоги»* → *«Один из осужденных полицейских - Николай Плебух - свою вину в суде не признал, считая, что действовал правомерно. Сообщалось, что он, находясь в нетрезвом состоянии, повредил радиатор батареи в камере для задержанных и получил сильные ожоги»*. В отрывке из первичного источника видно, что речь в тексте идет о двух разных людях. Последнее предложение говорится о том, что ожоги получил правонарушитель, который находился в камере, однако в текст реферата не вошло среднее предложение о правонарушителе, вследствие чего получилось так, будто ожоги получил осужденный полицейский.

Подводя итоги анализа, можно сделать вывод о том, что большего всего ошибок по всем параметрам в совокупности было обнаружено у первого алгоритма. Всего ошибок по всем критериям насчитывается 127. У второго алгоритма цифра менее внушительная – 95.

При сравнении двух алгоритмов, можно заключить, что в первом алгоритме допускалось больше ошибок в плане структуры, содержания и полноты представления информации. Данные ошибки обусловлены тем, что первый алгоритм сильнее ужимал первичные тексты, вследствие чего в текст реферата попадало меньшее количество предложений. Второй алгоритм менее подвержен такого рода ошибкам, так как при реферировании текстов рефераты получаются не такими сжатыми, как в случае с первым алгоритмом. Таким образом, можно сделать вывод, что в плане отражения структуры, содержания и полноты представления информации второй алгоритм работает лучше первого, однако текст реферата второго алгоритма получается достаточно большим, что позволяет избегать ошибок, которым подвержен первый алгоритм, но при этом нарушается функция реферата – кратко и лаконично доносить основную суть первоисточника.

Оба анализируемых алгоритма практически в равной степени подвержены ошибкам, которые связаны с разрывом логики и связностью текста. Ошибка нарушения логики и связности текста зачастую связана с отсутствием лида и неверным определением границ предложения.

Второй алгоритм в отличие от первого в большей мере подвержен ошибке наличия повторов. Данная ошибка связана с тем, что при реферировании в текст реферата второго алгоритма попадает больше предложений, чем в реферате первого алгоритма. Таким образом, первым алгоритмом исключаются излишние повторы в тексте реферата.

Второй алгоритм проявил себя хуже первого в плане наличия ошибок, связанных с искажением информации. Данный тип ошибки возникает из-за некорректной экстракции предложений, которые выстраивают текст реферата.

Общее количество хороших рефератов в первой и второй выборках практически одинаково. В первой выборке было насчитано 7 хороших рефератов из 100, а во второй выборке было насчитано 8 хороших рефератов из 100. При этом новостные тексты, рефераты которых могут считаться хорошими, в выборках не пересекаются, за исключением одного новостного текста. Другими словами, хорошие рефераты первого алгоритма в большинстве своем не являются таковыми в выборке рефератов второго алгоритма. С примерами хороших рефератов можно ознакомиться в Приложении Б (см. Приложение Б).

Таким образом, можно подытожить, что оба алгоритма нуждаются в существенной доработке, так как в обеих выборках более 90% текстов на выходе имеют те или иные ошибки. Сложно сказать, какой из алгоритмов сработал лучше, так как результаты анализа

показывают, что у обоих алгоритмов разные «слабые» места, однако они имеются и требуют устранения.

Выводы по второй главе

На основе вышеизложенного можно сделать вывод о том, что область автоматического реферирования нуждается в дополнительных исследованиях, направленных на изучение и разработку алгоритмов, способных автоматически реферировать тексты. Несмотря на то, что в обществе активно обсуждается проблема информационной перенасыщенности, алгоритмы автоматического реферирования найти не так просто.

Кроме того, как показал анализ, проведенный в рамках данной исследовательской работы, существующие алгоритмы, которые имеются в открытом доступе, нуждаются в существенной доработке, так как их рефераты в большинстве своем получаются некорректными.

Так, у анализируемых алгоритмов получилось практически одинаковое количество хороших рефератов. В первой выборке, которая выполнялась алгоритмов Text Compactor, насчитывается 7 хороших рефератов из 100. Во второй выборке, которая была собрана с помощью применения программного кода на языке программирования Python, насчитывается 8 хороших рефератов из 100. Такой показатель нельзя считать хорошим, так как более 90% рефератов имеют ошибки, что доказывает необходимость в дополнительных исследованиях в области автоматического реферирования.

В общем, исследуемые алгоритмы показали разную работу. Больше всего ошибок по всем критериям оказалось у первого алгоритма – 127, у второго алгоритма – 95.

Сравнение рефератов двух алгоритмов показало, что в первом алгоритме допускалось больше ошибок в структуре, содержании и полноте представления информации. Данным ошибкам есть разумное объяснение. Первый алгоритм сильнее ужимает тексты первоисточника, поэтому в реферат попадает меньше предложений, что влечет за собой вышеперечисленные ошибки.

В обоих алгоритмах насчитывалось практически равное количество ошибок, связанных с разрывом логики и связностью текста – 40 ошибок в первой выборке и 37 ошибок во второй выборке.

Как показал анализ, во второй выборке чаще встречается ошибка наличия повторов. Данная ошибка объясняется тем, что в текст реферата второго алгоритма входит больше предложений, чем в текст реферата первого алгоритма, что влечет за собой повтор информации.

Кроме того, в второй выборке насчитывается больше ошибок, связанных с искажением информации. Во второй выборке было насчитано 3 таких ошибки, а в первой – 1. Несмотря на то, что искажение информации встречалось не так часто, эта ошибка считается самой серьезной из всех возможных.

ЗАКЛЮЧЕНИЕ

С ростом количества информации в современном обществе растет и информационная перенасыщенность людей. Анализ огромного массива данных требует существенных человеческих ресурсов. Одним из способов автоматизации процесса обработки информации является автоматическое реферирование.

В рамках данной исследовательской работы мы фокусируемся на реферировании новостных текстов, так как именно новостные тексты являются ключевым источником прироста данных в сети Интернет. Кроме того, фокус на новостных текстах обусловлен заинтересованностью в автоматическом реферировании со стороны новостных порталов и агентств.

На сегодняшний день разрешение проблемы автоматического реферирования является актуальной задачей в области компьютерной лингвистики. Проблема автоматического реферирования заключается не только в создании алгоритма, который будет качественно реферировать тексты, но и в разработке системы оценивания качества автоматических рефератов, так как без системы оценивания невозможно понять актуальна ли на данный момент проблема разработки качественного алгоритма.

В ходе данного исследования мы занимались решением второй проблемы – разработка системы оценивания качества автоматических рефератов. Полученные результаты анализа подтвердили необходимость создания алгоритма, рефераты которого будут соответствовать критериям качества. Оба анализируемых алгоритмов показали достаточно низкий уровень качества рефератов. Так, более 90% текстов из обеих выборок содержали в себе ошибки.

В рамках проведения анализа, на этапе подготовки возникли трудности с поиском работающих алгоритмов автоматического реферирования, так как большинство онлайн рефераторов более не являются действующими, либо работают некорректно, а программные коды, находящиеся в свободном доступе при запуске выдают ошибки. Данная трудность еще раз указывает на то, что область автоматического реферирования нуждается в свежих разработках или доработке уже существующих алгоритмов.

Таким образом, при работе над исследованием была достигнута поставленная цель – разработать систему оценки эффективности алгоритмов реферирования, примененных к материалу новостных текстов. Проведен анализ автоматических рефератов новостных текстов, который указал на то, что область автоматического реферирования нуждается в совершенствовании алгоритмов.

Перспективы дальнейшего исследования нам видятся в разработке системы оценки эффективности алгоритмов автоматического реферирования на материале других типов текстов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Абрамова, Н. Н. Автоматическое составление обзорных рефератов новостных сюжетов / Н. Н. Абрамова, В. Е. Абрамов // Интернет-математика 2007 : сб. работ участников конкурса науч. проектов по информ. поиску / [отв. ред. П. И. Браславский]. — Екатеринбург : Изд-во Урал. ун-та, 2007. — С. 5–15.
2. Андреева К.А. Проектирование автоматизированной системы реферирования текстов / К.А. Андреева, Р.С. Шайдуров // Решетневские чтения. — 2015. — №19. — С. 202 — 203. — URL: <https://cyberleninka.ru/article/n/proektirovanie-avtomatizirovannoy-sistemy-referirovaniya-tekstov> (дата обращения: 28.02.2021).
3. Батура Т. В. Методы и системы автоматического реферирования текстов : монография / Т. В. Батура, А. М. Бакиева. Ин-т систем информатики им. А. П. Ершова СО РАН. — Новосибирск : ИПЦ НГУ, 2019. — 110 с.
4. Блюменау Д. И. Проблемы свертывания научной информации. Л. : Наука, 1982. — 166 с.
5. Вейзе А.А. Реферирование текста / А.А. Вейзе. — Минск : Изд-во Белорус. гос. ун-та, 1978. — 128 с.
6. Википедия: Свободная Энциклопедия [Электронный ресурс] // Автоматическое реферирование. — Электрон. дан. — [Б. м.], — 2020. — URL: https://ru.wikipedia.org/wiki/%D0%90%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D1%80%D0%B5%D1%84%D0%B5%D1%80%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5 (дата обращения: 12.02.2021).
7. ГОСТ 7.76-96. Библиографирование. Каталогизация. Основные понятия. URL: <https://files.stroyinf.ru/Data2/1/4294850/4294850382.pdf> (дата обращения 15.04.2021).
8. ГОСТ 7.9-95. Реферат и аннотация. Общие требования. URL: <http://www.tehnorma.ru/> (дата обращения: 24.01.2021).
9. Гречихин А.А. Жанры информационной литературы: Обзор. Реферат / А.А. Гречихин, И.Г. Здоров, В.И. Соловьев. — М. : Книга, 1983. — 320 с.
10. Губанова Е. Г. Рекомендации по написанию рефератов: для обучающихся по специальности 29.02.04 «Конструирование, моделирование и технология швейных изделий». — Магнитогорск: ГБОУ ПОО МТК, 2017. — 16 с.
11. Дейк Т. А. ван. Когнитивные и речевые стратегии выражения этнических предубеждений // Язык. Познание. Коммуникация. — М. : Прогресс, 1989. — С. 123.

12. Добросклонская Т.Г. Вопросы изучения медиатекстов. Опыт исследования современной английской медиаречи / Т.Г. Добросклонская. — М. : Ridero, 2005. — 170 с.
13. Добросклонская Т.Г. Теория и методы медиалингвистики / Т.Г. Добросклонская. — М. : АДД, 2000. — 203 с.
14. Догма газеты «Ведомости» [Электронный ресурс]. — Электрон. дан. — [Б. м.], — 2001. — URL: https://www.dropbox.com/s/7fof0qv7ps3qzg8/dogma_vedomosti.pdf?dl=0 (дата обращения: 10.05.2021).
15. Жолкова А. И. Фасетный метод реферирования как проблема библиографирования технической литературы: автореф. дис. ... канд. пед. наук. Л.: ЛГИК им. Н.К. Крупской, 1985. — 15 с.
16. Информационное агентство «ТАСС» [Электронный ресурс] // Рубрика: происшествия. — Электрон. дан. — [Б. с.], 2020. URL: <https://tass.ru/> (дата обращения: 14.05.21).
17. Казак М. Ю. Специфика современного медиатекста. В кн.: Г. Я. Солганик, Н. И. Клушина, В. В. Славкин и др. (ред.). Лингвистика речи. Медиастилистика: коллективная монография, посвященная 80-летию проф. Г. Я. Солганика. — М. : Флинта; Наука, 2012. — С. 320 — 334.
18. Казак М.Ю. Современные медиатексты: проблемы идентификации, делимитации, типологии / М.Ю. Казак // Медиалингвистика. — 2014. — № 1(4). — С.65— 76.
19. Колесникова Н.И. От конспекта к диссертации: учеб. пособие по развитию навыков письменной речи / Н.И Колесникова. — М. : Флинта: Наука, 2008. — 288 с.
20. Кривенко Б. В. Язык массовой коммуникации: лексико-семиотический аспект / Б.В. Кривенко. — Воронеж: Изд-во Воронежского университета, — 1993. — 136 с.
21. Лингвистический энциклопедический словарь. М., 1990. — С. 507.
22. Лингвистический энциклопедический словарь. М., 1990. — С. 136.
23. Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание. — 2009. — Выпуск 8 (15). — С. 299 — 305.
24. Маркушевская Л.П. Аннотирование и реферирование (Методические рекомендации для самостоятельной работы студентов) / Л.П. Маркушевская, Ю.А. Цапаева. — СПб ГУ ИТМО, 2008. — 51 с.

25. Нестерова Н. М. Реферативный перевод: проблема смыслового свертывания и семантической адекватности / Н.М. Нестерова // Вестник ЧелГУ. Сер. Филология. Искусствоведение. — 2011. — Вып. 58, №25(240). — С. 112 — 118.
26. Нестерова Н.М. Реферирование как способ извлечения и представления основного содержания текста / Н.М. Нестерова, Н.А. Герте // Вестник Пермского университета. Российская и зарубежная филология. — 2013. — №4 (24). — URL: <https://cyberleninka.ru/article/n/referirovanie-kak-sposob-izvlecheniya-i-predstavleniya-osnovnogo-soderzhaniya-teksta> (дата обращения: 25.02.2021).
27. Новиков А.И., Нестерова Н. М. Реферативный перевод научно-технических текстов. М. : Ин-т языкозн. АН СССР, 1991. — 148 с.
28. Османова М.М. Автоматическое реферирование и аннотирование текста / М.М. Османова, З.Г. Алиев // Инновационные аспекты развития науки и техники. — 2020. — №3. — URL: <https://cyberleninka.ru/article/n/avtomaticheskoe-referirovanie-i-annotirovanie-teksta> (дата обращения: 20.05.2021).
29. Подготовка и редактирование информационных изданий [Электронный ресурс]. — Электрон. дан. — М. : Центр дистанцион. образования МГУП, 2001. — URL: http://www.hi-edu.ru/e-books/PodgRedactInflzd/podg_inf_izd011.htm (дата обращения: 12.02.2021).
30. Реферирование научного текста: Методические рекомендации по курсу «Русский язык и культура речи» / Казан. гос. ун-т; Филол. фак-т; Каф. совр. рус. яз.; Сост.: А.В. Бастриков, Е.М. Бастрикова. — Казань : Казан. гос. ун-т, 2005. — 28 с.
31. Рябуха О. В. К вопросу типологии и интерпретации медиатекстов. Исследования языка и современное гуманитарное знание. — 2020. — Т. 2, № 2, — С. 72–80.
32. Солганик Г.Я. К определению понятий "текст" и "медиатекст" // Вестник Московского университета. Серия 10. Журналистика. — 2005. — №2. — С. 7 — 13. — URL: <https://cyberleninka.ru/article/n/k-opredeleniyu-ponyatiy-tekst-i-mediatekst> (дата обращения: 24.03.2021).
33. Соловьев В.И. Поаспектный метод реферирования / В.И. Соловьев // НТИ. Сер. 2. — 1971. — №2. — С.14 — 17.
34. Тарасов С. Д. Современные методы автоматического реферирования / С.Д. Тарасов // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика, телекоммуникации и управление. — 2010. — №6 (113). — С. 59 — 72. — URL: <https://cyberleninka.ru/article/n/sovremennyye-metody-avtomaticheskogo-referirovaniya> (дата обращения: 28.02.2021).

35. Тарасов С.Д. Автоматическое сводное реферирование новостных сообщений / С.Д. Тарасов // Научно-технический вестник информационных технологий, механики и оптики. — 2010. — № №4 (68). — С.88 — 93. — URL: <https://cyberleninka.ru/article/n/avtomaticheskoe-svodnoe-referirovanie-novostnyh-soobscheniy> (дата обращения: 28.02.2021).
36. Тарасов С.Д. Автоматическое составление обзорных рефератов новостных сюжетов / С.Д. Тарасов // Вестник Балтийского государственного технического университета. — 2008. — № 3. — С. 61— 67
37. Тревгода, С.А. Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: автореф. дис. ... канд. техн. наук / С.А. Тревгода. — СПб., 2009. — 18 с.
38. Чернявская В.Е. Текст в медиальном пространстве / В.Е. Чернявская. — М. : URSS; Либроком, 2013. — 228 с.
39. Шевченко И. Суммаризация с помощью TextRank. — [Электронный ресурс] // Блог. — Электрон. данн. — [Б. м.], 2013. — URL: <http://igorshvchenko.ru/blog/entries/textrank> (дата обращения: 20.12.21).
40. Шереметьева С. О. Интерактивное реферирование, ориентированное на машинный перевод / С.О. Шереметьева // Вестник ЮУрГУ. Серия: Лингвистика. — 2013. — №1. — С. 89 — 92. — URL: <https://cyberleninka.ru/article/n/interaktivnoe-referirovanie-orientirovannoe-na-mashinnyu-perevod> (дата обращения: 28.02.2021).
41. Школа журналистики [Электронный ресурс] // Информационные жанры. Как написать новость? — Электрон. дан. — [Б. м.], — 2017. — URL: <https://xn---7sbafuabraerjymxvsmn8f.xn--p1ai/informacionnye-zhanry-kak-napisat-novost/> (дата обращения: 12.04.2021).
42. Шмелева Т. В. Медиалингвистика: терминологический аспект / Т.В. Шмелева // Вестник Новгородского государственного университета им. Ярослава Мудрого. — 2015. — № 4-1 (87). — С. 32 — 35.
43. Яцко В.А. Симметричное реферирование: теоретические основы и методика / В.А. Яцко // НТИ. Сер. 2. — 2002. — № 5. — С. 18 — 28.
44. Analyticsvidhya [Electronic source] // An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation). — Electronic data. [W. p.], 2018. — URL: <https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/>

45. Autosummarizer [Electronic source] // Automatic Text Summarizer Multi-Language Summarizing & Paraphrasing AI Tool. — Electronic data. [W. p.], 2021. — URL: <http://autosummarizer.com/>
46. Bell A. Approaches to Media Discourse / A. Bell, P. Garrett. — Blackwell: Wiley, 1998. — 304 p.
47. Bell A. The language of the news media / A. Bell. — Oxford : Blackwell Publ., — 1991. — 277 p.
48. Cleveland D. Introduction to Indexing and Abstracting / D. Cleveland, A. Cleveland. — Littleton: Libraries Unlimited, 1983. — 283 p.
49. Corner J. Documentary television: the scope for media linguistics / J. Corner // AILA Review. — 1995. — № 12 : Applied Linguistics Across Disciplines. — P. 62 — 63. — URL: <http://www.aila.info/download/publications/review/AILA12.pdf> (access date: 14.01.2021).
50. Crystal D. Language and the Internet. 2nd ed. / D. Crystal. — Cambridge : Cambridge University Press, — 2006. — 304 p.
51. Dorr B., Monz C., President S., et al. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?, ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, ACL, Ann Arbor, MI. — 2005. — P. 1–8.
52. Edmundson H.P. New methods in automatic extraction // Journal of the Association for Computing Machinery. — 1969. Vol. 16, №. 2. — P. 264 — 285.
53. Edmundson H.P., Wyllys R.E. Automatic abstracting and indexing – survey and recommendations // Communications ACM. — 1961. Vol. 4, № 5. — P. 226 — 234.
54. Esummarizer [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <http://esummarizer.com/main/summarize>
55. Fairclough N. Language and power / N. Fairclough. — London : Longman Group Ltd, — 1989. — 259 p.
56. Fowler R. Language in the news: Discourse and ideology in the press / R. Fowler. — London: Routledge Publ., 1991. — 254 p.
57. GeeksforGeeks [Electronic source] // Python. Text Summarizer. — Electronic data. [W. p.], 2020. — URL: <https://www.geeksforgeeks.org/python-text-summarizer/>
58. Greatsummary [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <http://www.greatsummary.com/?f>
59. Helpfulpapers [Electronic source] // Free Text Summary Generator. — Electronic data. [W. p.], 2021. — URL: <https://helpfulpapers.com/summarizer-tool>

60. Jensen K. Mass Communication Research Methods / K. Jensen, N. Jankowski. — New York: Routledge, 1991. — 268 p.
61. Jurafsky D. Speech and Language Processing [Electronic source] / Jurafsky D., Martin J. H. — 2018. — URL: <https://web.stanford.edu/~jurafsky/slp3/> (access date: 05.01.2020).
62. Loret E. A Gradual Combination of Features for Building Automatic Summarisation Systems Text / E. Lloret, M. Palomar // Speech and Dialogue. — Heidelberg, 2009. — P. 16 — 23
63. Luhn H. The automatic creation of literature abstracts // IBM Journal of Research and Development. — 1958. Vol. 2, № 2. — P. 159 — 165.
64. McKeown K., Passonneau R.J., Elson D.K., et al. Do summaries help? A task-based evaluation of multi-document summarization, 28th International Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil, ACM. — 2005. — P. 210 — 217.
65. Medium [Electronic source] // Automatic Text Summarization Made Simple with Python. — Electronic data. [W. p.], 2020. — URL: <https://medium.com/luisfredgs/automatic-text-summarization-made-simple-with-python-f9c3c645e34a>
66. Medium [Electronic source] // Summarise Text with TFIDF in Python. — Electronic data. [W. p.], 2018. — URL: <https://medium.com/@shivangisareen/summarise-text-with-tfidf-in-python-bc7ca10d3284>
67. Mihalcea R. TextRank: Bringing Order into Text / R. Mihalcea, P. Tarau // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. — Barcelona : Association for Computational Linguistics. — 2004. — P. 404 — 411.
68. Montgomery M. An introduction to language and society / M. Montgomery. — London; New York : Methuen Publ., 1986. — 211 p.
69. Oxford English Dictionary [Electronic source] / Oxford University Press. — Electronic data. — [W. p.], — 2021. — URL: <https://www.oed.com/> [access data: 20.03.2021]
70. Prateek J. An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation). [Electronic source] // Analytics Vidhya — Electronic data. — [W. p.], 2018. URL: <https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/> (access date: 16.09.2020).
71. Sakai T. Generic summaries for indexing in information retrieval / T. Sakai, K. Spark-Jones // ACM Special Interest Group on Information Retrieval (SIGIR'01): 24th International Conference on Research and Development in Information Retrieval, New Orleans, LA, ACM. — 2001 pp. 190 — 198.

72. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings / K. Bennani-Smires, C. Musat, A. Hossmann et al. // Proceedings of the 22nd Conference on Computational Natural Language Learning. — Brussels : Association for Computational Linguistics, 2018. — P. 221 — 229. — URL: <https://www.aclweb.org/anthology/K18-1022.pdf> (access date: 16.09.2020).
73. Smmry [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <https://smmry.com/>
74. Spärck Jones K., Galliers J.R. Evaluating Natural Language Processing Systems: An Analysis and Review // LNCS. — 1996. Vol. 1083. — P. 26 — 30.
75. Sumplify [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <https://www.awesomedomains.com/sale/sumplify.com>
76. Swesum [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <http://swesum.nada.kth.se/index-eng.html>
77. Text Compactor [Электронный ресурс]. — Электрон. дан. — [Б. м.], — 2006 — 2010. — URL: <http://www.textcompactor.com/> (дата обращения: 10.05.2021).
78. Textteaser [Electronic source]. — Electronic data. [W. p.], 2021. — URL: <http://textteaser.com/>
79. Tools4noobs [Electronic source] // Online summarize tool (free summarizing). — Electronic data. [W. p.], 2021. — URL: <https://www.tools4noobs.com/summarize/>
80. Torres-Moreno J.M. Automatic Text Summarization / J.M. Torres-Moreno. — London : ISTE ; Hoboken, NJ : Wiley, 2014. — 376 p.
81. Towards data science [Electronic source] // Process Text using TFIDF in Python. — Electronic data. — [W. p.], 2018. — URL: <https://towardsdatascience.com/tfidf-for-piece-of-text-in-python-43feccaa74f8>
82. Traditional methods for text data [Electronic resource] // Towards data science. — Electronic data. — [W. p.], 2018. — URL: <https://towardsdatascience.com/understanding-feature-engineering-part-3-traditional-methods-for-text-data-f6f7d70acd41> (access date: 05.04.2020).
83. Van Dijk T. Text and Context: Explorations in the Semantics and Pragmatics of Discourse / T. Van Dijk . — London : Longman, 1980. — 282 p.
84. Wikipedia: The Free Encyclopedia [Electronic source] // MS Fnd in a Lbry. — Electronic data. — [W. p.], — 2021. — URL: https://en.wikipedia.org/wiki/MS_Fnd_in_a_Lbry (access data: 12.04.2021).

ПРИЛОЖЕНИЕ А

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
from nltk.corpus import stopwords
stopWords = set(stopwords.words("russian"))
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize, sent_tokenize

text_str = ''

...
```

Рисунок 1 – Необходимые для работы библиотеки и ввод данных.

```
def _create_frequency_table(text_string) -> dict:
    words = word_tokenize(text_string)
    stemmer = PorterStemmer()

    freqTable = dict()
    for word in words:
        word = stemmer.stem(word)
        if word in stopWords:
            continue
        if word in freqTable:
            freqTable[word] += 1
        else:
            freqTable[word] = 1
    return freqTable
```

Рисунок 2 – Создание таблицы частотности слов.

```

def _score_sentences(sentences, freqTable) -> dict:
    sentenceValue = dict()
    for sentence in sentences:
        word_count_in_sentence = (len(word_tokenize(sentence)))
        word_count_in_sentence_except_stop_words = 0
        for wordValue in freqTable:
            if wordValue in sentence.lower():
                word_count_in_sentence_except_stop_words += 1
            if sentence[:10] in sentenceValue:
                sentenceValue[sentence[:10]] += freqTable[wordValue]
            else:
                sentenceValue[sentence[:10]] = freqTable[wordValue]
        if sentence[:10] in sentenceValue:
            sentenceValue[sentence[:10]] = sentenceValue[sentence[:10]]
            / word_count_in_sentence_except_stop_words

    return sentenceValue

```

Рисунок 3 – Присвоение оценки предложениям.

```

def _find_average_score(sentenceValue) -> dict:

    sumValues = 0
    for entry in sentenceValue:
        sumValues += sentenceValue[entry]

    average = (sumValues / len(sentenceValue))

    return average

```

Рисунок 4 – Поиск средней оценки предложений.

```

def _generate_summary(sentences, sentenceValue, treshold):
    sentence_count = 0
    summary = ''

    for sentence in sentences:
        if sentence[:10] in sentenceValue and sentenceValue[sentence[:10]] >= (treshold):
            summary += " " + sentence
            sentence_count += 1

    return summary

```

Рисунок 5 – Сохранение предложений в реферат.

```

if __name__ == '__main__':
    result = run_summarization(text_str)
    print(result)

```

Рисунок 6 – Вывод реферата.

ПРИЛОЖЕНИЕ Б

Таблица 1 – Примеры хороших рефератов первого алгоритма

Текст первоисточника	Текст реферата
Сенатор Олег Королев прокомментировал ДТП в Липецке По словам политика, инцидент произошел из-за проблем с сердцем. Он отрицает, что находился в нетрезвом состоянии МОСКВА, 14 мая. /ТАСС/. Российский сенатор, бывший губернатор Липецкой области Олег Королев заявил в пятницу ТАСС, что ДТП в Липецке с участием его автомобиля произошло из-за проблем с сердцем. "Просто у меня был резкий рост кровяного давления вместе с мерцательной аритмией. Я просто съехал с дороги, искал место, где мне съехать, напоролся на что-то, быстро отъехал. Я позвонил сотрудникам ГАИ", - пояснил Королев. Сенатор отрицает, что в момент ДТП находился в нетрезвом состоянии. "У нас сейчас в социальных сетях что угодно могут написать. Пусть пишут, на чужой роток не накинешь платок. Я абсолютно трезв", - добавил Королев. Сенатор уточнил, что уже доложил в Совет Федерации о произошедшем.	Сенатор Олег Королев прокомментировал ДТП в Липецке По словам политика, инцидент произошел из-за проблем с сердцем. Российский сенатор, бывший губернатор Липецкой области Олег Королев заявил в пятницу ТАСС, что ДТП в Липецке с участием его автомобиля произошло из-за проблем с сердцем. "Просто у меня был резкий рост кровяного давления вместе с мерцательной аритмией. Сенатор уточнил, что уже доложил в Совет Федерации о произошедшем. Ранее Telegram-канал Baza со ссылкой на видео и фотографии из соцсетей сообщил, что автомобиль Lexus, на котором якобы часто передвигается Королев, сбил дорожный знак. Как пояснил ТАСС источник в правоохранительных органах, по факту ДТП в Липецке с участием Королева составлен протокол об административном правонарушении. Королев назначен членом Совета Федерации от Липецкой области с ноября 2018 года, после того как президент РФ Владимир Путин принял его досрочную отставку с должности губернатора.

Ранее <u>Telegram-канал</u> <u>Baza</u> со ссылкой на видео и фотографии из соцсетей сообщил, что автомобиль Lexus, на котором якобы часто передвигается Королев, сбил дорожный знак. Как пояснил ТАСС источник в правоохранительных органах, по факту ДТП в Липецке с участием Королева составлен протокол об административном правонарушении. Королев назначен членом Совета Федерации от Липецкой области с ноября 2018 года, после того как президент РФ Владимир Путин принял его досрочную отставку с должности губернатора. Он руководил регионом с 1998 года.	
Текст первоисточника	Текст реферата
Информация о пропавшем воздушном судне в Сибири не подтвердилась В точку, где откуда поступил сигнал аварийного радиомаяка, вылетал вертолет, поиск результатов не дал НОВОСИБИРСК, 14 мая. /ТАСС/. Спасатели проверили <u>информацию</u> о срабатывании аварийного маяка воздушного судна в Минусинском районе Красноярского края, в данном районе ничего не обнаружено, сообщили ТАСС в Сибирском авиационно-спасательном центре. "Сработал аварийный радиомаяк. В данную точку полетел Ми-8, ничего там не	Информация о пропавшем воздушном судне в Сибири не подтвердилась В точку, где откуда поступил сигнал аварийного радиомаяка, вылетал вертолет, поиск результатов не дал НОВОСИБИРСК, 14 мая. Спасатели проверили информацию о срабатывании аварийного маяка воздушного судна в Минусинском районе Красноярского края, в данном районе ничего не обнаружено, сообщили ТАСС в Сибирском авиационно-спасательном центре. "Сработал аварийный радиомаяк. Поиск результатов не дал", - сказал собеседник агентства. В Западно-Сибирском следственном

обнаружил, вернулся. Поиск результатов не дал", - сказал собеседник агентства.

В Западно-Сибирском следственном управлении на транспорте СК РФ уточнили ТАСС, что аварийный радиомаяк неизвестного воздушного судна подал сигнал бедствия в Хакасии, и он не зарегистрирован в международной поисково-спасательной спутниковой системе.

"В результате обследования данного участка местности какого-либо воздушного судна не обнаружено. Радиомаяк, который подавал сигнал бедствия, не зарегистрирован в системе КОСПАС-САРСАТ (международная спутниковая поисково-спасательная система - прим. ТАСС)", - говорится в сообщении.

Следственные органы Западно-Сибирского следственного управления на транспорте СК России проводят по данному факту доследственную проверку. Уточняется, что инцидент произошел в 11:50 (07:50 мск) в 40 км от Абакана.

Ранее сообщалось, что в Минусинском районе Красноярского края зафиксирован сигнал срабатывания аварийного радиомаяка.

По данным муниципальных властей, поиски велись у села Городок - населенного пункта на берегу Тубинского залива Енисея. В месте поисков зафиксировано возгорание. В пресс-службе краевого главка МЧС сообщили журналистам, что

управлении на транспорте СК РФ уточнили ТАСС, что аварийный радиомаяк неизвестного воздушного судна подал сигнал бедствия в Хакасии, и он не зарегистрирован в международной поисково-спасательной спутниковой системе.

"В результате обследования данного участка местности какого-либо воздушного судна не обнаружено. Уточняется, что инцидент произошел в 11:50 (07:50 мск) в 40 км от Абакана. Ранее сообщалось, что в Минусинском районе Красноярского края зафиксирован сигнал срабатывания аварийного радиомаяка.

тлело возгорание на свалке, которое уже потушено.	
---	--

Таблица 2 – Примеры хороших рефератов второго алгоритма

Текст первоисточника	Текст реферата
Площадь возгорания строящегося детсада на Кубани увеличилась более чем в три раза Сейчас она составляет 1 000 кв. м КРАСНОДАР, 14 мая. /ТАСС/. Площадь <u>возгорания</u> строящегося детского сада в Кореновске увеличилась более чем в три раза - до 1 000 кв. м, по факту случившегося организовано проведение проверки, сообщили журналистам в прокуратуре Краснодарского края. ГУ МЧС ранее сообщало, что в Кореновске загорелся строящийся детский сад, площадь пожара составляет 300 кв. м. По данным оперативного штаба региона, пострадавших нет. "По уточненным данным, площадь возгорания составляет не менее 1 000 кв. м, общая площадь комплекса - 12 000 кв. м. Органы прокуратуры по Краснодарскому краю организовали проверку по данному факту. Установлено, что возгорание произошло на чердачном помещении при завершении монтажных работах, предварительно установлено, [что причиной могло стать] неосторожное обращение с огнем. На место выехал	Площадь возгорания строящегося детсада на Кубани увеличилась более чем в три раза Сейчас она составляет 1 000 кв. м КРАСНОДАР, 14 мая. /ТАСС/. Площадь возгорания строящегося детского сада в Кореновске увеличилась более чем в три раза - до 1 000 кв. м, по факту случившегося организовано проведение проверки, сообщили журналистам в прокуратуре Краснодарского края. По данным оперативного штаба региона, пострадавших нет. "По уточненным данным, площадь возгорания составляет не менее 1 000 кв. м, общая площадь комплекса - 12 000 кв. На место выехал заместитель прокурора края Михаил Леговец", - заявили в пресс-службе. В ведомстве добавили, что прокуратура проверит соблюдение норм пожарной безопасности при проведении строительных работ.

заместитель прокурора края Михаил Леговец", - заявили в пресс-службе. В ведомстве добавили, что прокуратура проверит соблюдение норм пожарной безопасности при проведении строительных работ.	
Текст первоисточника	Текст реферата
После отравления в колледже Хабаровского края к медикам обратились 20 человек В медучреждениях остаются 12 человек ХАБАРОВСК, 14 мая. /ТАСС/. Медики оказали помощь 19 учащимся и одному сотруднику колледжа в Комсомольске-на-Амуре, поступившими к ним с признаками отравления за два дня. Госпитализированы 12 человек, об этом сообщили журналистам в пресс-службе Министерства здравоохранения Хабаровского края. Ранее прокуратура Хабаровского края сообщила, что в инфекционные отделения больниц Амурска и поселка Солнечный с отравлением помещены более десяти студентов строительного колледжа Комсомольска-на-Амуре. Следователи возбудили по этому факту уголовное дело по ч. 1 ст. 238 УК РФ ("Оказание услуг, не отвечающих требованиям безопасности жизни и здоровья потребителей"). "Всего за два дня [13 и 14 мая] были госпитализированы 20 человек. Так как	После отравления в колледже Хабаровского края к медикам обратились 20 человек В медучреждениях остаются 12 человек ХАБАРОВСК, 14 мая. /ТАСС/. Медики оказали помощь 19 учащимся и одному сотруднику колледжа в Комсомольске-на-Амуре, поступившими к ним с признаками отравления за два дня. Госпитализированы 12 человек, об этом сообщили журналистам в пресс-службе Министерства здравоохранения Хабаровского края. Следователи возбудили по этому факту уголовное дело по ч. 1 ст. 238 УК РФ ("Оказание услуг, не отвечающих требованиям безопасности жизни и здоровья потребителей"). "Всего за два дня [13 и 14 мая] были госпитализированы 20 человек. При поступлении состояние у всех пациентов оценивалось как средней тяжести. Сейчас в больницах остаются 12 человек. Расследованием причин отравления занимаются специалисты Роспотребнадзора.

инфекционное отделение городской больницы № 2 г. Комсомольска-на-Амуре по-прежнему работает в режиме ковида, госпитализация производилась в инфекционные отделения Солнечной (11 человек) и Амурской ЦРБ (девять человек)", - сказали в пресс-службе. При поступлении состояние у всех пациентов оценивалось как средней тяжести. Сейчас в больницах остаются 12 человек. Расследованием причин отравления занимаются специалисты Роспотребнадзора.	
--	--

ИНФОРМАЦИЯ О ДОКУМЕНТЕ

№ документа: 48
Начало загрузки: 19.06.2021 12:02:21
Длительность загрузки: 00:00:01
Имя исходного файла: Диссер_проверка на антиплагиат.txt
Название документа: Диссер_проверка на антиплагиат
Размер текста: 144 кБ
Символов в тексте: 147095
Слов в тексте: 18297
Число предложений: 1624

ИНФОРМАЦИЯ ОБ ОТЧЕТЕ

Начало проверки: 19.06.2021 12:02:23
Длительность проверки: 00:00:18
Комментарии: не указано
Модули поиска: Интернет



Заимствования — доля всех найденных текстовых пересечений, за исключением тех, которые система отнесла к цитированиям, по отношению к общему объему документа.

Самоцитирования — доля фрагментов текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника, автором или соавтором которого является автор проверяемого документа, по отношению к общему объему документа.

Цитирования — доля текстовых пересечений, которые не являются авторскими, но система посчитала их использование корректным, по отношению к общему объему документа. Сюда относятся оформленные по ГОСТу цитаты; общеупотребительные выражения; фрагменты текста, найденные в источниках из коллекций нормативно-правовой документации.

Текстовое пересечение — фрагмент текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника.

Источник — документ, проиндексированный в системе и содержащийся в модуле поиска, по которому проводится проверка.

Оригинальность — доля фрагментов текста проверяемого документа, не обнаруженных ни в одном источнике, по которым шла проверка, по отношению к общему объему документа.

Заимствования, самоцитирования, цитирования и оригинальность являются отдельными показателями и в сумме дают 100%, что соответствует всему тексту проверяемого документа.

Обращаем Ваше внимание, что система находит текстовые пересечения проверяемого документа с проиндексированными в системе текстовыми источниками. При этом система является вспомогательным инструментом, определение корректности и правомерности заимствований или цитирований, а также авторства текстовых фрагментов проверяемого документа остается в компетенции проверяющего.

№	Доля в отчете	Доля в тексте	Источник	Актуален на	Модуль поиска	Блоков в отчете	Блоков в тексте
[01]	3,55%	3,55%	Скачать книгу Т.Г. Добросклонской «Медиалингвистика: системный подход к изучению языка СМИ» (1/2) http://ffl.msu.ru	10 Дек 2018	Интернет	51	51
[02]	3,27%	3,27%	Нестерова Н. М., Герте Н. А. (Россия, Пермь) http://rfp.psu.ru	17 Окт 2014	Интернет	39	39
[03]	0,85%	0,85%	Методические рекомендации по написанию рефератов Учебно-методическое пособие для обучающихся по профессии 262019.02 Закройщик и специальности 262019 Конструирование, моделирование и	10 Фев 2017	Интернет	3	3