Федеральное государственное бюджетное учреждение науки
Институт системного анализа Российской академии наук

На правах рукописи

Швец Александр Валерьевич

ВЗАИМОДЕЙСТВИЕ ИНФОРМАЦИОННЫХ И ЛИНГВИСТИЧЕСКИХ МЕТОДОВ В ЗАДАЧАХ АНАЛИЗА КАЧЕСТВА НАУЧНЫХ ТЕКСТОВ

Специальность 05.13.17 – Теоретические основы информатики

**Диссертация на соискание ученой степени
кандидата технических наук**

Научный руководитель:
доктор физико-математических наук,
профессор Геннадий Семенович Осипов

Москва 2015

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. В открытой научной печати регулярно появляются тексты, которые не прошли должную проверку перед изданием. Они не соответствуют требованиям, предъявляемым к научным публикациям, содержат различные нарушения или вовсе являются псевдонаучными. Такие тексты встречаются в научных журналах (обычно не рецензируемых), в материалах конференций и в других источниках научной сферы (под источниками научной сферы понимаются издания открытой печати и информационные ресурсы, которые позиционируют себя как научные). В большинстве случаев нарушения приводят к снижению ясности изложения, что вводит в заблуждение как исследователей, которые знакомятся с новой для них научной областью, так и аналитиков, работающих с большими объемами данных, у которых нет возможности рассматривать каждый текст детально. Существующие методы автоматического анализа текстов не ориентированы на проверку качества анализируемых текстов. Они позволяют выполнять поиск релевантных запросу публикаций, структурировать данные, извлекать полезную информацию, однако отсутствие этапа, на котором определяется надежность источника и возможность использования содержащейся в нем информации, часто приводит к некорректным, необъективным результатам. В связи с этим требуется разработка методов и программных средств автоматического определения признаков, характеризующих качество текстов научной сферы, и выявления псевдонаучных текстов. Под качеством понимается совокупность характеристик, включающих оценку лексики и множества синтактико-семантических структур текста, оценку наличия лингвистических ошибок, оценку наличия псевдонаучных фрагментов, оценку формальной структуры текста, т. е. наличия в тексте необходимых разделов (например, описания результатов). Настоящая работа посвящена созданию методов интеллектуального анализа текстов, направленных на решение указанных задач, что свидетельствует о ее актуальности.

Извлечение признаков, характеризующих качество текста, опирается на лингвистические структуры, выделяемые в тексте посредством синтаксического и семантического анализа, а также на информационные методы: статистические, индуктивного порождения гипотез и машинного обучения. Множество признаков формируется на основе лексических, морфологических, синтаксических и информационных характеристик текстов научной сферы.

Научная задача. Разработка нового метода определения качества текстов научной сферы, основанного на автоматическом выявлении лексических, синтаксических, морфологических и информационных признаков.

Предмет исследования – методы автоматического обнаружения признаков, характеризующих качество текстов научной сферы.

Основной целью является автоматизация определения качества текстов научной сферы. Для достижения поставленной цели в работе решаются следующие **задачи**:

1. Выполнить анализ методов определения различных характеристик и свойств текстов научной сферы.
2. Разработать метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
3. Разработать метод автоматического выявления структуры научной публикации.
4. Разработать метод автоматического обнаружения лингвистических ошибок.
5. Разработать метод автоматического определения псевдонаучных фрагментов текстов научной сферы.
6. Сформировать признаковое пространство для автоматического определения научных и псевдонаучных текстов.
7. Проверить экспериментально разработанные методы.

Методы исследования. В диссертации использованы методы интеллектуального анализа текстов, статистические методы, методы машинного обучения, методы снижения размерности признакового пространства, индуктивные методы порождения гипотез, метод реляционно-ситуационного анализа текстов.

Научная новизна и результаты, выносимые на защиту.

1. Разработан новый метод автоматического формирования общенаучного словаря устойчивых словосочетаний.
2. Разработан новый метод автоматического выявления структуры научной публикации.
3. Разработан новый метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.
4. Впервые разработан метод автоматического выявления псевдонаучных фрагментов текстов научной сферы.
5. Сформировано множество признаков, характеризующих качество текстов научной сферы.
6. Построено множество правил для обнаружения псевдонаучных текстов.

Теоретическая значимость работы состоит в создании новых методов автоматического выявления признаков, характеризующих качество текстов научной сферы, на основе взаимодействия информационных и лингвистических методов.

Практическая значимость. Результаты работы могут применяться в системах поддержки принятия решений при отборе заявок, проектов, приеме отчетов, статей для публикации в научных журналах и в трудах конференций, а также для решения иных задач интеллектуального анализа информации. Разработанные методы извлечения признаков научного текста и метод обнаружения псевдонаучных текстов могут применяться в системах поиска и анализа научной информации.

Реализация результатов работы. Разработанные методы определения качества текстов научной сферы реализованы в виде программных средств и внедрены в следующие организации:

- Государственная публичная научно-техническая библиотека (информационная система «ЭКБСОН»);
- ООО «Национальный цифровой ресурс «Руконт» (электронно-библиотечная система «Руконт»);
- ООО «Научно-издательский центр ИНФРА-М» (электронно-библиотечная система «Znanium.com»);
- ЗАО «РосИнтернет технологии» (система интеллектуального поиска и анализа научных публикаций «Exactus Expert»).

Разработанные методы, правила и алгоритмы использованы в рамках научно-исследовательских работ по следующим проектам Минобрнауки РФ, программам ОНИТ РАН и грантам РФФИ:

1. Создание программного комплекса информационно-аналитической поддержки научно-технической деятельности на основе вычислительного семантического поиска и анализа неструктурированной текстовой информации (*ФЦП, № 07.551.11.4003, 2011-2013 гг.*);
2. Разработка вычислительных методов объективной оценки качества научно-технических документов на естественных языках (*ФЦП, № 14.514.11.4018, 2012-2013 гг.*);
3. Исследование и разработка методов и алгоритмов связанности сложно-структурированных данных в научно-технической сфере (*ФЦП, № 14.514.11.4024, 2012-2013 гг.*);
4. Развитие методов и технологии семантического поиска и анализа научных публикаций Exactus Expert (*в рамках проекта 2.9 ОНИТ РАН 2012-2013 гг.*);

5. Исследование методов и разработка моделей и средств оценки научных текстов на основе их когнитивных структур (*грант РФФИ № 14-29-05028-офи_м, 2014-2016 гг.*).

Достоверность результатов подтверждена проведенными вычислительными экспериментальными исследованиями программных средств, реализующих предложенные методы, правила и алгоритмы.

Апробация результатов исследования. Основные положения работы докладывались и обсуждались на следующих научных конференциях:

- XVI Международная научная конференция «Решетневские чтения», ноябрь 2012, г. Красноярск.
- Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012, октябрь 2012, г. Белгород.
- Пятая международная конференция «Системный анализ и информационные технологии» (САИТ-2013), сентябрь 2013, г. Красноярск.
- 20-я Международная конференция "Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса", июнь 2013, г. Судак.
- III Всероссийская научная конференция молодых ученых с международным участием «Теория и практика системного анализа» (ТПСА’14), май 2014, г. Рыбинск.
- Восемнадцатая международная научно-практическая конференция "SCIENCE ONLINE: электронные информационные ресурсы для науки и образования", май 2014, г. Белек.

– 7-я Международная конференция «Интеллектуальные системы» IEEE (The 7th IEEE International Conference Intelligent Systems, IS’2014 IEEE, Warsaw), сентябрь 2014, г. Варшава.

– Шестая международная конференция «Системный анализ и информационные технологии» (САИТ-2015), июнь 2015, г. Светлогорск.

Публикации. По теме диссертации опубликовано 9 работ [1-9], из них 4 в рецензируемых изданиях, рекомендованных ВАК РФ и приравненных к ним [1-4], и 2 зарегистрированные программные системы [5, 6]. Опубликованные в этих работах результаты, относящиеся к методам и алгоритмам выявления признаков, характеризующих качество текстов научной сферы, получены лично автором.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения, списка использованных источников и приложения. В приложении приведены описания программ, реализующих алгоритмы, предложенные в работе. Работа изложена на 120 страницах машинописного текста, содержит 21 таблицу и 24 рисунка. Список использованных источников включает 94 наименования.

В **первой главе** рассматриваются нарушения в текстах научной сферы, которые могут приводить к снижению ясности изложения текста и к отсутствию возможности оценить исследование, применить описанные методы и воспроизвести эксперименты. В первом параграфе приводится типология нарушений в научных публикациях и исследуется возможность их выявления с помощью анализа лексики и синтаксических структур. Во втором параграфе рассматриваются методы, позволяющие в некоторой степени выполнять автоматический анализ качества научных текстов. В заключительной части главы приведены основные выводы и сформулированы задачи исследования.

Вторая глава посвящена разработке методов выделения признаков, характеризующих качество текстов научной сферы, а именно разработке

метода автоматического формирования общенаучного словаря устойчивых словосочетаний, метода автоматического выявления структурных разделов научной публикации и метода автоматического обнаружения лингвистических ошибок. Выполнено экспериментальное исследование методов.

В **третьей главе** выполняется исследование применимости разработанных методов для выявления псевдонаучных текстов. В первом параграфе приводится определение псевдонауки, используемое в настоящей работе. Второй параграф посвящен разработке метода автоматического определения псевдонаучных фрагментов. В третьем и четвертом параграфах формируется признаковое пространство и выполняется индуктивное построение множества правил для обнаружения псевдонаучных текстов. В заключительной части приводится сравнение методов машинного обучения, подходящих для решения задачи классификации текстов научной сферы с целью обнаружения псевдонаучных текстов на основе сформированного пространства признаков.

В **заключении** приводятся основные результаты, полученные в работе.

В **приложении** описаны реализованные программные модули, которые внедрены в программный комплекс интеллектуального поиска и анализа научных публикаций «Exactus Expert» и использованы при тестировании разработанных методов. Приведены снимки системы и примеры отчетов, получаемых в результате работы программных модулей.

Глава 1. ИССЛЕДОВАНИЕ НАРУШЕНИЙ В ТЕКСТАХ НАУЧНОЙ СФЕРЫ

1.1. Типология нарушений в научных публикациях

Объектом исследования настоящей работы является множество текстов научной сферы. Среди них можно выделить как научные тексты, так и тексты, которые претендуют на то, чтобы быть научными, но содержат различные нарушения принципов научного исследования, которые делают текст незаконченным, малоинформативным или вовсе ненаучным.

Основные цели научной публикации – закрепление процесса познания и хранение знания, сообщение информации и доказательство ее истинности, – определяют характерные особенности научного стиля. Согласно [10], основная задача научной и технической литературы – предельно ясно и точно довести определенную информацию до читателей, что достигается логически обоснованным изложением фактического материала. Создавая научный текст, автор имеет возможность тщательно продумать композицию, отобрать наиболее точные слова и однозначные грамматические конструкции, удовлетворяющие требованиям, предъявляемым к качеству изложения [11]. Однако установка на определенное качество речи не всегда дает желаемый результат, исследования показывают, что современная научная литература наполнена речевыми погрешностями и другими нарушениями, затрудняющими понимание содержания [11, 12].

Среди множества нарушений в текстах научной сферы можно выделить следующие типы:

- Нарушение требований к лексике научного текста;
- Нарушение структуры научного текста;
- Нарушение правил согласования;
- Нарушение синтаксической и семантической связности;
- Лексическая избыточность (употребление плеоназмов);
- Нарушение последовательности изложения.

Рассмотрим последовательно различные типы нарушений, изучив требования к текстам научных статей и примеры, когда эти требования не выполняются, и исследуем, могут ли такие тексты быть выявлены путем анализа лексики и синтаксических структур. В ходе исследования необходимо определить, какие признаки, соответствующие нарушениям, характеризуют качество текстов научной сферы.

1.1.1. Нарушение требований к лексике научного текста

При написании научных текстов принято использовать научный функциональный стиль речи, который обладает следующими свойствами: обобщенно-отвлеченный характер речи, подчеркнутая логичность, последовательность изложения, его доказательность и аргументированность, точность, ясность, объективность, некатегоричность изложения. Выражение этих свойств в тексте происходит в основном на уровне лексики, морфологии и синтаксиса [13, 14]. Точность научного стиля достигается употреблением большого числа терминов, абстрактной лексики и устойчивых общенаучных словосочетаний. В [15] под общенаучными словосочетаниями и выражениями понимают научные и технические термины и различные выражения общего характера, такие как: *принятая гипотеза, по указанной причине, обосновать вывод, описанный ниже* и другие. Использование таких выражений позволяет логично выстроить содержание текста, передать мысль наиболее строгим образом. Отклонение от научного стиля приводит к снижению ясности изложения, часто начинает прослеживаться субъективный характер высказываний.

Рассмотрим в качестве примера фрагмент публикации, автор которой использует просторечную лексику, что нарушает свойство отвлеченности научного стиля.

Пример 1: *«И что об этом думают сами языковеды? Не стану добавлять имеющуюся словесную чепуху с целью придания наукообразия ссылками на разнообразные мнения на сей счет. Их без труда можно найти в*

Интернете. Из всех предлагаемых толкований ясно одно: происхождение и смысл слова "смерд" авторам не известны. Некоторые даже включают его в словарь иностранных слов. То есть не русских. О чем это говорит?»

Анализ фрагмента показывает, что использованная лексика не соответствует принятым требованиям к научному стилю речи. Видно, что в нем отсутствуют синтаксические структуры характерные для научных текстов, и присутствуют обычно не используемые словосочетания, такие как: *«думают языковеды», «словесная чепуха», «придание наукообразия», «найти без труда»* и другие.

Рассмотрим другой пример научного текста, который содержит небольшой процент общенаучных словосочетаний и написан в стиле близком к разговорному.

Пример 2: *«Какой ты станешь теперь, Россия? Трудно предугадать. Тем не менее в насущных поисках незаемного, обретаемого через страдания пути одоления хаоса - к возрождению и процветанию не должна быть отбрасываема неотъемлемая исторически для нашей самобытности проблематика взаимодействия культур России и Запада. Причем рассматриваемая не с одной лишь негативной стороны, как нередко теперь бывает, - разумеется, небеспричинно. А и с плодотворной. Для этого надобно настроить себя также на то, чтобы скорее оставить в прошлом все, так или иначе делавшее Россию в продолжение десятилетий культурным гетто».*

Наличие в представленном фрагменте таких синтаксических конструкций как *«трудно предугадать», «в насущных поисках», «нередко бывает», «надобно настроить»* делает текст более красочным, однако, для того чтобы сохранялась ясность и объективность проведенного исследования, требуются более точные формулировки.

Так, текст с низким употреблением общенаучных словосочетаний и высокой долей неупотребимых в языке словосочетаний становится менее понятным, неточным и, соответственно, менее информативным. Текст, в котором совсем не используются устойчивые общенаучные словосочетания,

как правило, не является научным. В связи с этим в качестве признаков, характеризующих качество научного текста, может быть выбрано количество устойчивых общенаучных словосочетаний в тексте и количество неупотребимых словосочетаний. Номинальными значениями первого признака могут быть следующие: «приемлемое», «заниженное», «низкое». Значениями второго признака могут быть: «низкое», «завышенное», «высокое». Для определения значений признаков требуется общенаучный словарь устойчивых словосочетаний, словарь сочетаемости слов языка и методы автоматического установления в текстах синтаксических связей.

Рассмотрим следующий тип нарушений и соответствующие ему признаки, характеризующие качество текста.

1.1.2. Нарушение структуры научного текста

Научное сообщество вырабатывает стандарты, которые призваны оптимизировать процессы распространения передовых идей и представлений, способствовать повышению информированности и возможности полезного взаимодействия ученых. Основным средством обмена информацией между учеными являются первичные научные тексты (первое публичное представление существенной информации о проведенном исследовании). К ним относятся публикации в научных журналах, отчеты о научно-исследовательской работе и прочие тексты, представляющие новые результаты научной деятельности. Для того чтобы одни ученые могли повторно использовать результаты, полученные другими учеными, при написании первичных научных текстов необходимо учитывать общепринятые требования к структуре публикации.

Согласно [16], приемлемая первичная научная публикация представляется в форме, которая позволяет коллегам оценить исследование, воспроизвести эксперименты и оценить интеллектуальный процесс, приведший к выводам. Такой текст содержит информацию о предмете, методах, целях и результатах научного исследования, проведённого в соответствии с

методологическими принципами объективности и системности. Структура качественного научного текста, как правило, соответствует формату IMRAD [17] (Introduction, Methods, Results, and Discussion – IMRAD), согласно которому статья, описывающая результаты оригинального экспериментального исследования, должна включать в себя следующие основные разделы: «Постановка проблемы», «Методы», «Результаты» и «Выводы». Если статья посвящена теоретическому исследованию, то раздел Methods заменяется на Theoretical Basis [17].

Научные публикации в формате IMRAD впервые появились в конце XIX века [18]. В настоящее время этот формат стал универсальным стандартом, принятым большинством журналов. В 1972 г. Национальный американский институт стандартов одобрил и рекомендовал IMRAD для применения, что определило дальнейшее распространение тенденции к унификации структуры публикаций, посвященных результатам оригинальных исследований. В англоязычной периодике уже к 1970-м гг. доля оформленных в соответствии с IMRAD статей составляла 80%, а начиная с 1980-х гг. тексты с отличающейся структурой к публикации не принимаются [19]. Большинство современных российских научных журналов предъявляют идентичные требования к статьям [18].

Приведем примеры синтаксических структур, характерных для отдельных разделов. В разделе «Постановка проблемы», как правило, используются следующие словосочетания: *«поставлена задача», «поиск средства», «проведение анализа», «один из подходов», «необходимость изучения», «приобретает актуальность»* и другие. Для раздела «Методы» характерны следующие выражения: *«анализировать состав», «методика заключается в», «последующее измерение», «определять по методу», «характеристика выборки»* и другие. Остальные разделы также имеют специальные конструкции и речевые обороты.

Полное отсутствие в научном тексте лексики и синтаксических конструкций, свойственных некоторому структурному разделу, будет говорить

об отсутствии этого раздела, что свидетельствует о нарушении структуры научного текста. В связи с этим оценки наличия каждого раздела в отдельности могут служить еще одним признаком качества научных текстов. Их получение, как и в случае с определением количества устойчивых общенаучных словосочетаний, может быть основано на применении лингвистического анализа текста и исследовании лексики и синтаксических структур.

1.1.3. Нарушение правил согласования

В потоке публикаций немалую долю составляют тексты, в которых обнаруживаются те или иные отступления от норм научного изложения, выражающиеся в виде лингвистических ошибок. Одной из распространенных ошибок, встречающихся в научных текстах, является отсутствие согласования различных частей речи. Можно выделить следующие виды нарушений согласования:

- Нарушения согласования прилагательных с существительными в роде, числе и падеже;
- Нарушения подчинительной связи прилагательного;
- Нарушения согласования сказуемого с однородными подлежащими;
- Нарушения согласования причастия с определяемыми словами, стоящими перед причастным оборотом;
- Неоднозначность связи причастий с определяемыми словами в причастных оборотах;
- Неправильное употребление превосходной степени прилагательного.

Приведем несколько примеров предложений из научных статей, содержащих нарушения согласования. Курсивом выделены несогласованные слова.

Пример 3: «Такие факторы как *возраст*, *образование*, социальный *статус* обычно *оказывает* существенное влияние на речевое поведение носителя языка».

Пример 4: «На выходе блока 5 управления формируется сигнал запроса на ввод и код N_{3U}, *пропорциональный напряжения* U_3, записывается в вычислительный блок 4».

Пример 5: «Как правило, этот параязыковой прием сопровождается … *паузами* хетизации, *присущих* языковым личностям с высоким уровнем притязаний…».

Пример 6: «Еще одна *особенность* социальной символизации, *связанной* с употреблением собеседниками ненормативной лексики…».

Такие нарушения можно обнаружить, построив правила, которые будут использовать синтаксический разбор предложения и морфологический разбор слов. Наличие нарушений согласования является еще одним признаком низкого качества текста.

1.1.4. Нарушение синтаксической и семантической связности

Из рассмотренных выше примеров видно, что в некоторых случаях, формально, синтаксические связи могут быть установлены между несвязными словами, как в примере 6 между словами «символизации» и «связанной». Однако часто слова остаются без связей и отделяются от синтаксического дерева, как причастие «присущих» в примере 5. Большое число таких случаев в тексте будет говорить о низкой синтаксической связности текста. Близким нарушением является низкая семантическая связность текста, обычно выражающаяся в неправильном глагольном управлении.

Приведем примеры предложений, встречающиеся в научной литературе, с нарушением семантических связей.

Пример 7: «Сформулировать и *доказать о свойствах* прямоугольных треугольников».

В примере 7 ошибочно использовано дополнение в предложном падеже с предлогом «*о»* при глаголе «доказать» [20]. Такая ошибка является частой и возникает под влиянием сочетаний типа: «подумать о чём-либо», «рассказать о чём-либо».

Пример 8: «Эти работы, опубликованные уже почти полвека тому назад, *опирались на результатах* исследований, выполненных к тому времени».

В примере 8 допущено неправильное управление при глаголе «опираться». В данном случае вместо предложного падежа должен быть употреблен винительный падеж – «опираться на результаты» [20].

Пример 9: «Использование синонимов в речи помогает *избежать повторение* одних и тех же слов».

В примере 9 выбран неправильный падеж при глаголе «избежать», который требует дополнения в родительном падеже [20].

Выделенные в примерах ошибки могут быть выявлены путем анализа синтаксических и семантических структур. Наличие в тексте подобных нарушений может стать еще одним признаком, характеризующим качество научного текста.

1.1.5. Лексическая избыточность

Другим типом ошибок, затрудняющих понимание содержания, является нарушение норм лексической стилистики, в частности, лексическая избыточность – неоправданное многословие, которое встречается в научных текстах в виде так называемых плеоназмов.

Под плеоназмом понимается дублирование некоторого элемента смысла; наличие нескольких языковых форм, выражающих одно и то же значение, в пределах законченного отрезка речи или текста – а также само языковое выражение, в котором имеется подобное дублирование [21]. Плеоназм принято подразделять на обязательный, т.е. обусловленный языковой системой, и факультативный, т.е. не обусловленный языковой системой; факультативные плеоназмы бывают конвенциональные (закрепленные языковой нормой) и

неконвенциональные, т.е. создаваемые заново говорящим или пишущим [22]. Последний вид плеоназмов часто встречается в научных текстах низкого качества, поэтому степень употребления плеоназмов может служить еще одним признаком, определяющим качество научного текста.

Приведем несколько примеров неоправданного многословия из научных текстов и текстов, позиционирующих себя как научные. Курсивом выделены повторяющиеся лексемы.

Пример 10: «Материальное тело, двигаясь в любом направлении, всегда создаст результирующий темп *времени* из своего темпа *времени* и движения *времени* от расширения или сжатия шара *времени*».

Пример 11: «Высший предмет *иерархии* – компонент, называющий высший предмет в *иерархии* в моделях, выражающих отношения *иерархии*/каузацию отношений *иерархии* (подчиняться закону)».

Пример 12: «Протяженность *поля* внутренней сферы, так же как и протяженность *поля* ядра находятся в определенной зависимости от величины плотности первичного *поля*, но в результате однородности *полей*, плотность *поля* внутренней сферы находится в зависимости от силовой характеристики гравитационного *поля* ядра Земли».

Как видно, прочтение предложений, представленных в примерах, занимает много времени, что может отвлекать от понимания содержания всего текста. Рассмотрим еще один тип нарушения, заключающегося в отклонении от последовательности изложения.

1.1.6. Нарушение последовательности изложения

Для обозначения порядка явлений и связей между ними используются парные элементы, указывающие на последовательность изложения. В качестве таких элементов могут выступать, например, словосочетания «с одной стороны, с другой (стороны)», вводные слова «во-первых, во-вторых» и другие [23]. Отсутствие одного из этих элементов при наличии другого является нарушением в тексте.

Приведем несколько примеров, соответствующих этому типу нарушений.

Пример 13: «Алгоритм помогает снять многозначность *двумя способами*: *во-первых*, он снижает количество возможных значений пересечением…; *если* пересечение *всё же* содержит более одного значения, то из оставшихся можно выбрать…».

Отсутствие в примере 13 вводного слова «во-вторых» приводит к тому, что при беглом прочтении обнаруживается только один способ «снятия многозначности». Можно догадаться, что со слов «*если всё же*» начинается описание второго способа, но с первого взгляда, кажется, что это лишь описание дополнительных условий к первому способу. Таким образом, опущенное вводное слово приводит к неоднозначности толкования высказывания.

Следующий пример показывает, что незавершенность фраз может быть признаком необоснованности высказываний. Не закончив одну мысль, автор сменяет ее другой, что приводит к пустословию, сбивающему с толку читателя, и к неправильным выводам.

Пример 14: «Судьбе было угодно, чтобы закон Хаббла … вобрал в себя максимум противоречий. *Во-первых*, почему собственно закон Хаббла это закон. Хаббл выявил всего лишь не очень явную зависимость…. В любом случае, закономерность, выявленная Хабблом, является гипотетической… Таким образом, гипотетическая закономерность, выявленная Хабблом, была трансформирована в закон, который непостижимым образом, без всякой проверки, был включен во все справочники».

Рассмотренная типология нарушений покрывает часто встречающиеся ошибки в текстах научной сферы. При этом из приведенных примеров следует, что такие ошибки характеризуются лексикой и синтаксическими структурами, которые могут быть выделены автоматически с помощью методов обработки естественного языка. Систематизация рассмотренных признаков, которые характеризуют качество текстов научной сферы, приведена на рис. 1.

Рисунок 1 – Признаки, характеризующие качество текстов научной сферы

В следующем разделе представлен обзор современных методов, позволяющих в некоторой степени выполнять автоматический анализ качества научных текстов.

1.2. Методы автоматического анализа качества научных текстов

Существует большое число работ, описывающих, как правильно подготовить научную статью [24, 25] и как оценить качество научного исследования вручную [26, 27]. Автоматическое оценивание публикаций в основном выполняется посредством анализа косвенных показателей цитирования с помощью различных наукометрических методов [28]. Недостатком такого подхода является возможность оценивания только публикаций, попавших в цитатные базы, в которые, как правило, включено небольшое число высокорейтинговых журналов. К тому же возможность

оценить новую публикацию появляется спустя длительное время, которое требуется на ее изучение научным сообществом, написание и издание работ с ее цитированием и индексацией этих работ цитатными базами. Такие ограничения наукометрических методов для оценки публикаций приводят к необходимости разработки методов, позволяющих в реальном времени автоматически извлекать из текста признаки, которые могут характеризовать его качество. Рассмотрим задачи, возникающие при автоматическом анализе качества текстов научной сферы, и современные исследования по их решению.

Одна из задач автоматического анализа текста связана с исследованием лексики: анализ текста может позволить определить написан ли текст понятным для исследователей языком с использованием устойчивых общенаучных словосочетаний. В [15] предложен метод формирования словаря общенаучных выражений с использованием разработанного языка LSPL (LexicoSyntactic Pattern Language) для записи лексико-синтаксических шаблонов [29]. В рамках этого языка шаблонов составлялся словарь маркеров операций научного дискурса. Материалом для словаря служили общедоступные словари словосочетаний научной речи [30, 31]. Они анализировались вручную, и по некоторым условиям отбирались те слова и словосочетания, для которых организационная функция в научном дискурсе является очевидной. Полнота построенного словаря не определялась, по мнению авторов, требуется его некоторая доработка. Разработанные лексико-синтаксические шаблоны и составленные словари маркеров операций научного дискурса использовались для распознавания логико-композиционной структуры научного текста. Показателей эффективности предлагаемого метода авторы не приводят.

Другая задача состоит в определении информативности текста. Так, ведется разработка методов, позволяющих определить, состоит ли текст лишь из предложений общего характера или его содержание является узкоспециальным. Например, в [32] предлагается метод определения общих и специфичных фраз для автоматической оценки читабельности текста.

Утверждается, что хорошо написанный текст содержит в равном количестве, как предложения общего, обзорного характера, так и предложения специфичного характера, придающие высказываниям конкретику. Если текст содержит слишком много общих фраз, он является недостаточно информативным, и наоборот, излишне специфичный текст может ввести читателя в заблуждение. Для классификации фраз 5 экспертов оценили вручную 2800 примеров общих и специфичных предложений, которые были использованы для обучения линейного классификатора, основанного на методе логистической регрессии. Для предложений, на которых оценки всех экспертов совпали, классификатор смог предсказать правильный класс в 95% случаев; для предложений, на которых согласованность была только у четырех из пяти экспертов, точность составляла 85%. В случае, когда совпадали оценки лишь трех экспертов, точность классификации была равна 75%. Другой метод, предложенный в [33], определяет минимальное число тематик, требующееся для покрытия заданного процента информации, содержащейся в тексте. Этот метод используется в [34] для оценки информационной сжимаемости документа, что позволяет определить относится ли текст к одной-двум предметным областям или размыт по многим областям науки, что также является показателем информативности текста. Для тематического моделирования применяется метод латентно-семантического анализа [35], который работает с векторным представлением текстовых фрагментов типа «bag-of-words». Метод используется в [34] при попарном сравнении качества научных текстов. Эксперименты проведены на выборке малого размера (10 статей), что не позволяет судить об эффективности метода.

В Мичиганском университете проводились исследования по выявлению в текстах спекулятивных рассуждений, которые требуют проверки на истинность [36]. Такие рассуждения часто применяются при написании научных статей, как правило, экспериментального характера, например, в области биомедицины. Они содержат специальные языковые конструкции, используемые авторами, для того чтобы подчеркнуть неуверенность в

умозаключениях, правильность которых еще не доказана, но которые могут быть полезны в дальнейших исследованиях. Показано, что отдельные слова несут спекулятивный характер, но только в определенном контексте. Для выявления спекулятивных предложений построен список характерных ключевых слов, и предложено применить классификацию с помощью метода опорных векторов (SVM) с линейным ядром [37]. В качестве признаков использовались выделенные ключевые слова, их контекст, позиция предложения в тексте и некоторые лингвистические свойства слов и словосочетаний. Проводилась отдельно классификация предложений аннотаций и предложений полных текстов. Вычислены показатели эффективности классификации, такие как точность, полнота и F_1-мера [38]. Полученное значение F_1-меры для аннотаций составило 91,7%, что близко к верхней возможной границе, установленной авторами метода в результате анализа согласованности мнений экспертов, размечавших предложения для тестирования. Для предложений полных текстов значение F_1-меры было несколько ниже – 82,8%, что, по мнению авторов, связано с меньшим размером выборки (9 полных текстов статей и 1273 аннотации).

Некоторые современные исследования связаны с проблемой определения категорий предложений, соответствующих различным структурным разделам научной публикации. Эта проблема часто рассматривается как задача классификации. Например, в [39] использован мультиноминальный наивный байесовский классификатор [40] для распределения предложений биомедицинских статей по категориям «Введение», «Методы», «Результаты» и «Выводы». Исследованы такие признаки как слова, n-граммы, наличие цитирований, времена глаголов, позиция в тексте. Показано, что наибольшая точность классификации достигает 92% при учете всех признаков, и утверждается, что полученные категории предложений могут быть использованы при автоматическом извлечении полезной информации из текста. Другие исследователи предлагают выделять структурные категории для автоматического аннотирования полных текстов статей [41]. Авторы различают

11 категорий на уровне предложений: «Гипотеза», «Мотивация», «Цель», «Объект», «Фон», «Метод», «Эксперимент», «Модель», «Наблюдение», «Результат», «Вывод». На корпусе из 265 статей по химии и биохимии были обучены два классификатора, основанных на методе опорных векторов и методе условных случайных полей [42]. Были получены приемлемые результаты, однако с высокой точностью авторам удалось распознать лишь три категории предложений: «Эксперимент», «Фон» и «Модель». Для них значения F_1-меры составили 76%, 62%, and 53%, соответственно. Также этим коллективом исследователей предложен другой набор категорий: «Факт», «Гипотеза», «Интерпретация результатов», «Метод», «Проблема», «Цель» и «Результаты измерений» [43]. Разработанные методы использованы в рамках проекта ART-Project (ART-Project (2007-2009), SAPIENT Automation project (2009-2010)) при разработке инструмента SAPIENT (Semantic Annotation of Papers: Interface & ENrichment Tool) [44], позволяющего определять в научных текстах структурные категории предложений для автоматического построения аннотаций. Оба метода, предложенные в [39, 41] в качестве основного признака, значительно повышающего точность классификации, используют названия (заголовки) разделов анализируемой публикации. Это является их недостатком, поскольку применение методов ограничивается лишь несколькими предметными областями, у которых заголовки разделов соответствуют названиям структурных категорий предложений.

Выполненный обзор методов анализа текстов показывает, что признаки, характеризующие качество текста, могут быть получены автоматически. Однако существующие методы имеют недостатки и ограничения. Остается, например, нерешенной задача определения структуры научного текста, в случае, когда заголовки разделов не содержатся в тексте в явном виде. Также мало исследованы методы выявления нарушений требований к лексике, методы выявления лингвистических ошибок и ряда других нарушений.

1.3. Основные выводы и постановка задачи

Проведенное исследование различных нарушений в текстах научной сферы и выполненный обзор методов автоматического анализа качества научных текстов показывает возможность обнаружения нарушений с применением анализа лексики и синтаксических структур. Наличие или отсутствие в тексте определенного нарушения является признаком, характеризующим качество текста. Выделены следующие типы признаков:

- признаки, связанные с нарушением лексики;
- структурные признаки;
- признаки, связанные с лингвистическими ошибками;

Лексические, морфологические, синтаксические, семантические и информационные характеристики текста лежат в основе перечисленных выше признаков и могут быть получены с помощью методов глубокого лингвистического анализа и статистических методов. На рис. 2 представлена предлагаемая в настоящей работе схема извлечения рассмотренных признаков, характеризующих качество текста. Сначала происходит формирование базовых средств, а именно формирование общенаучного словаря устойчивых словосочетаний, выявление маркеров структурных разделов и формирование правил, характеризующих лингвистические ошибки. Затем выполняется анализ конкретного текста и извлечение его характеристик с помощью синтаксического и семантического анализа. После этого применяются методы выявления нарушений, которые оперируют с извлеченными характеристиками и сформированными базовыми средствами. В результате происходит формирование множества признаков, характеризующих качество анализируемого текста. Все методы, соответствующие процессам на рис. 2, за исключением синтактико-семантического анализа текста, разработаны в рамках настоящей работы и описаны в главе 2.

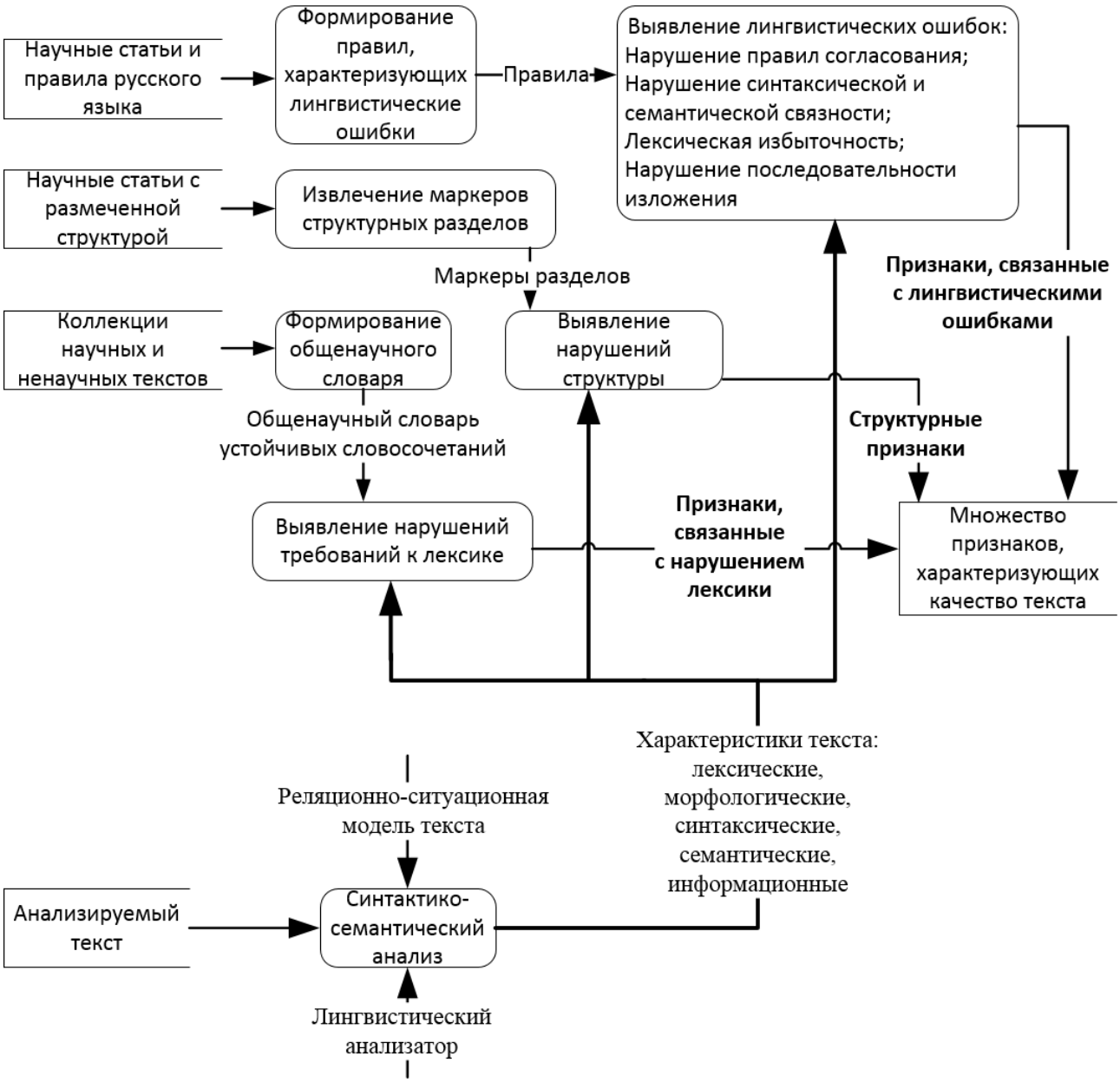


Рисунок 2 – Предлагаемая в работе схема выявления признаков, характеризующих качество текстов научной сферы

Для проверки применимости рассмотренных признаков для определения качества текстов научной сферы поставлена задача автоматического обнаружения псевдонаучных текстов, решение которой описывается в главе 3 настоящей работы.

Необходимость разработки методов автоматического извлечения представленных признаков и проверки их применимости определила следующие задачи:

1) разработать метод автоматического формирования общенаучного словаря устойчивых словосочетаний;
2) разработать метод автоматического выявления структуры научной публикации;
3) разработать метод обнаружения лингвистических ошибок, позволяющий выявлять нарушения синтаксической и семантической связности, лексическую избыточность, обнаруживать нарушения правил согласования, нарушения последовательности изложения;
4) разработать метод автоматического выявления псевдонаучных фрагментов;
5) сформировать признаковое пространство для автоматического определения качества текстов научной сферы;
6) исследовать применимость признаков для решения задачи автоматического обнаружения псевдонаучных текстов;
7) в рамках решения последней задачи построить множество правил, характеризующих принадлежность текста множеству псевдонаучных текстов;
8) провести сравнительный анализ эффективных методов классификации для обнаружения псевдонаучных текстов.

Решению задач 1-3 посвящена вторая глава, решение задач 4-8 рассмотрено в третьей главе.

Глава 2. МЕТОДЫ ВЫДЕЛЕНИЯ ПРИЗНАКОВ, ХАРАКТЕРИЗУЮЩИХ КАЧЕСТВО ТЕКСТОВ НАУЧНОЙ СФЕРЫ

Данная глава посвящена разработке методов выделения признаков, характеризующих качество текстов научной сферы, а именно разработке метода автоматического формирования общенаучного словаря устойчивых словосочетаний, метода автоматического выявления структурных разделов научной публикации и метода автоматического обнаружения лингвистических ошибок.

Особенностью всех предложенных методов является оперирование с полуструктурированными данными, которые формируются в результате синтактико-семантического анализа и представляют собой текст с установленными свойствами его элементов.

Перейдем к задаче формирования общенаучного словаря словосочетаний.

2.1. Выделение устойчивых общенаучных словосочетаний

Как отмечено в разделе 1.1.1, признак «количество устойчивых общенаучных словосочетаний» отвечает за качество научного текста, а для определения его значений требуется общенаучный словарь. Рассмотрим возможность применения существующих словарей для проверки соответствия текста научному стилю, выявим их достоинства и недостатки.

2.1.1. Словари общенаучной лексики

Формирование словарей общенаучной лексики обычно выполняется вручную с использованием общих словарей, среди которых – различные словари иностранных слов, энциклопедические словари, толковые словари и различные узконаправленные словари.

Некоторые общенаучные словари содержат слова, которые относятся скорее к общеупотребительным, чем к общенаучным. Например, слова «бюджет» и «вознаграждение» из словаря [45] без контекста могут относиться к

разным стилям речи (рис. 3), поэтому использование подобных словарей не позволит автоматически с большой точностью определить соответствие текста научному стилю.

Блоковые субсидии	Возмещение износа (основных средств)
Болонская декларация	Вознаграждение
Бритва Оккама	Возникновение авторского права
Брутто	Возраст оборудования
Бурение данных	Возрастная когорта
Бэнчмаркинг	Возрастной ценз
Бюджет	Вокабула
Бюджет времени	Воротнички (белые, серые, синие)
Бюджетный год	Восприятие
Бюллетень	Воспроизведение программы для ЭВМ или базы данных
Бюрократизация высшего образования	

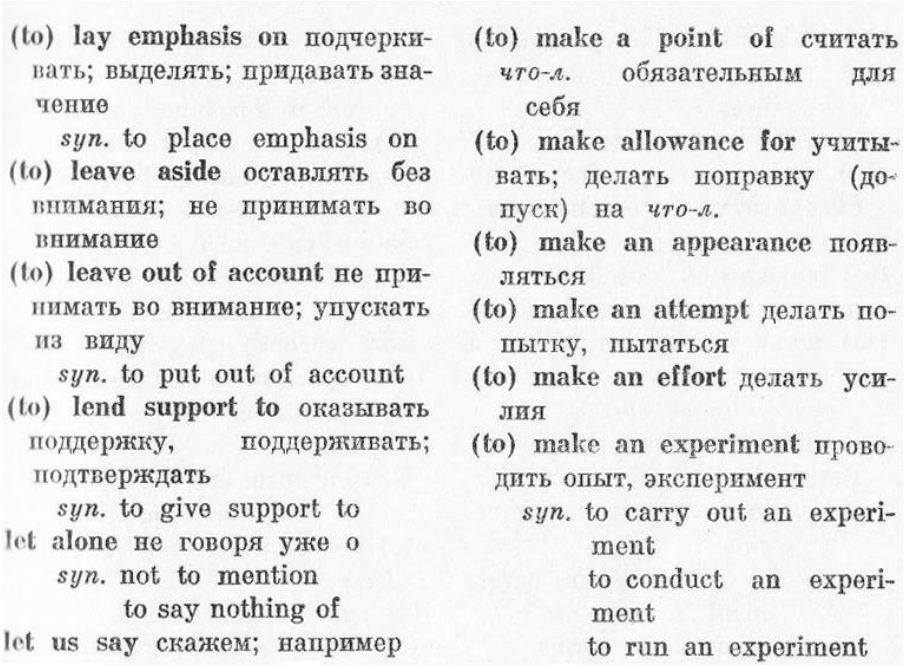
Рисунок 3 – Пример элементов словаря общенаучных слов и словосочетаний

Существуют сборники, представляющие типичные научные фразы для написания статей на английском языке [46, 31]. Они демонстрируют некоторые примеры, как можно сформулировать высказывание в научном стиле. Фрагменты таких сборников представлены на рис. 4 и рис. 5.

I. TEXT ORGANIZATION 16

- The general notion/ central concept/ (important) point ...in our model ...is that P.
  Our hypothesis/ The idea ...here ...is that P.
- This ...concept ...will serve as a starting point for studying P.
  ...notion ...needs careful explanation/ needs to be reconsidered.
  ...idea ...is a useful starting-point for investigating P.
  ...theory ...is basic/ fundamental/ strategic/ germane/ relevant to our approach.
  ...hypothesis ...proved useful in solving problems concerning P.
  ...view ...is prevailing/ innovative in contemporary research.
  ...interpretation ...can be generalized to apply to P.
  ...model ...can be extended to take account of P.
  The ...basic notions ...of this theory are given below/ in [NN].
- First, ...I am going to sketch my proposal that P/ to use/ study P as Q.
  ...I want to point to/ review/ introduce new principles.
  Now ...I can/ shall/ should provide P/ I must emphasize that P.
  ...I must say a few words about P.
  ...I should/ would like to illustrate/ suggest/ posit/ stress that P.
  Here ...we can make (put forth) a supposition that P.
  ...we propose a theory to account for P.
  At this point ...a certain clarification is necessary.
  ...one important detail must be noted, P.
  Let us now ...turn to/ return to example 1 (the discussion of P)/ (re)consider P.
  We must now...define/ determine/ establish/ estimate/ inquire into/ provide P.
  Finally, ...I find it necessary to consider P/ to turn our attention to P.
  The concluding/ final sections/ comments/ remarks concern/ focus on/ provide P.

Рисунок 4 – Примеры фраз из сборника «Научная речь на английском языке»



(to) lay emphasis on подчеркивать; выделять; придавать значение
syn. to place emphasis on
(to) leave aside оставлять без внимания; не принимать во внимание
(to) leave out of account не принимать во внимание; упускать из виду
syn. to put out of account
(to) lend support to оказывать поддержку, поддерживать; подтверждать
syn. to give support to
let alone не говоря уже о
syn. not to mention
to say nothing of
let us say скажем; например
(to) make a point of считать *что-л.* обязательным для себя
(to) make allowance for учитывать; делать поправку (допуск) на *что-л.*
(to) make an appearance появляться
(to) make an attempt делать попытку, пытаться
(to) make an effort делать усилия
(to) make an experiment проводить опыт, эксперимент
syn. to carry out an experiment
to conduct an experiment
to run an experiment

Рисунок 5 – Словарь глагольно-именных словосочетаний общенаучной речи

Как и другие построенные вручную общенаучные словари, подобные сборники, как правило, обладают высокой точностью, поскольку они сформированы компетентными экспертами, однако они имеют низкую полноту и не содержат многие словосочетания, встречающиеся в научных текстах. Автоматическое построение обобщений этих примеров для получения словаря всевозможных устойчивых словосочетаний является трудоемкой задачей, поскольку степень обобщения ограничивается семантически корректными сочетаниями слов.

В [15] приведен способ ручного формирования словаря общенаучных слов и словосочетаний, учитывающий семантику языка. Для разработки словаря были отобраны слова и словосочетания из доступных словарей словосочетаний научной речи [30, 31] и из научных статей различных областей науки (главным образом по компьютерным наукам и искусственному интеллекту). Отбор производился вручную с использованием следующих критериев:

– слово или словосочетание должно часто встречаться в текстах нескольких научных областей;

– организационная функция слова или словосочетания в научном дискурсе должна быть очевидна;
– семантически близкие словосочетания объединяются в группы как функционально эквивалентные.

Первое условие гарантирует, что входящие в словарь слова и словосочетания будут универсальными для всех или, по крайней мере, для большинства областей науки. Второе и третье условия связаны с целью построения исследователями такого словаря: решалась задача автоматического определения дискурсивной структуры научного текста. В [15] излагаются основные идеи распознавания структуры, однако показатели эффективности решения задачи с использованием такого словаря не приводятся, не определена его полнота. Утверждается, что словарь нуждается в тестировании и уточнении.

Основным недостатком перечисленных словарей является то, что представленные в них словосочетания по большей части являются именными группами и не учитывают семантические связи. Другой недостаток, связанный с ручным формированием, заключается в ограниченном, относительно небольшом числе примеров, на основе которых составляется словарь, что может сказываться на его полноте. Наконец, для того чтобы определить соответствие текста научному стилю, требуется коллекция не только общенаучных, но и собственно научных устойчивых словосочетаний, поскольку некоторые научные тексты написаны с минимальным использованием общенаучной лексики.

2.1.2. Установление синтаксических и семантических связей

В настоящей работе для формирования общенаучного словаря устойчивых словосочетаний предлагается выявлять в научных текстах словосочетания с синтаксическими и семантическими связями. Установление этих связей в тексте может быть выполнено с использованием лингвистических анализаторов, позволяющих автоматически строить деревья зависимостей и устанавливать значения синтаксем.

Методы, использующие синтаксический анализ, позволяют выполнять формирование словаря, в отличие от ручного составления, последовательно и объективно, с высокой полнотой (зависит от обрабатываемого корпуса) и большой скоростью. В [47] представлены эксперименты по составлению словаря на испанском языке, показано преимущество метода с синтаксическим анализом перед биграммным методом, которое выражается более высокими значениями точности и полноты. Для оценки использовались данные ручной разметки и результаты синтаксического разбора испанского текста. В [48] предлагается метод составления частотных словарей по корпусу текстов, который использует семантико-синтаксический анализ для выделения наименований понятий и дополнительный лингвистический анализ для исключения ошибочной, малоинформативной лексики. Утверждается [49], что такой метод автоматизации составления словарей позволяет в короткие сроки и с минимальными трудозатратами создать систему взаимосвязанных наименований понятий для заданной предметной области.

Существуют методы автоматического формирования словарей, которые не прибегают к синтаксическому анализу. Так, исследователи Института прикладной математики имени М.В. Келдыша Российской академии наук занимаются составлением словаря моделей управления глаголов, который используют для формирования словаря сочетаемости слов русского языка [50]. В своей работе авторы применяют несколько шаблонов, синтаксическая корректность которых, по мнению авторов, не вызывает сомнения в подавляющем большинстве случаев. Например, группа существительного, следующая за единственным глаголом в предложении, чаще всего синтаксически подчиняется данному глаголу, при этом прилагательные подчиняются существительному. Другим примером является конструкция «глагол + предлог + существительное», в которой обычно существительное подчиняется глаголу через предлог. Использование подобных шаблонов позволяет авторам избежать построения дерева зависимостей для всего предложения, что ускоряет процесс формирования словаря. В [51] также

предлагается уйти от полного анализа предложения и выполнять лишь частичный синтаксический анализ, устанавливая связи только для тех слов, для которых требуется извлечь информацию о сочетаемости. При таком подходе не выполняется построение дерева, включающего все слова исходного предложения, что позволяет существенно упростить алгоритм установления связей и ускорить его работу.

В настоящей работе методы поверхностного анализа не применяются, поскольку выполнение семантического анализа требует полного синтаксического разбора предложения. В связи с этим используется подход, основанный на построении синтаксических групп с помощью синтаксических правил, и его готовая реализация в системе АОТ [52].

Семантический анализ необходим для выделения в тексте и исследования конструкций, определяющих смысл текста. Они задаются словосочетанием, которое содержит семантическую связь. Такая конструкция в общем случае состоит из предиката и синтаксемы, которая замещает определенную роль [53]. Предикат обычно задается глаголом или отглагольным существительным; синтаксемой называется минимальная синтактико-семантическая единица языка, несущая обобщенный категориальный смысл и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков. Структурно-функциональные особенности синтаксем используются при построении компьютерных моделей многословных конструкций языка [54]. Набор значений синтаксем может быть произвольным и зависит от решаемой задачи. Показано [55], что семантический анализ позволяет формировать предметные знания, которые используются для интерпретации и отождествления извлекаемой информации.

Для установления значений синтаксем и выделения семантических конструкций в настоящей работе применяется метод реляционно-ситуационного анализа текстов, который опирается на словарь предикатных слов и основан на теории коммуникативной грамматики русского языка и теории неоднородных семантических сетей [53, 56].

Рассмотрим разработанный в рамках настоящей работы метод автоматического формирования общенаучного словаря словосочетаний.

2.1.3. Формирование общенаучного словаря устойчивых словосочетаний

Пусть S^+ и S^- – множества предложений научных и ненаучных текстов, таких что $\left|S^+\right| \le \left|S^-\right|$. Требуется построить словарь словосочетаний W, в большей степени характерных для предложений множества S^+. Предлагается следующая последовательность шагов.

Алгоритм 2.1 **(алгоритм формирования общенаучного словаря устойчивых словосочетаний).**

Шаг 1. Задать множества $W^+=\{\varnothing\}$ и $W^-=\{\varnothing\}$ – множества словосочетаний, входящих в предложения множеств S^+ и S^- соответственно.

Шаг 2. Выполнить синтактико-семантический разбор каждого предложения множества S^+, расширяя множество W^+ словосочетаниями с синтаксическими и семантическими связями.

Шаг 3. Для каждого встретившегося словосочетания w_i $(i=\overline{1,\left|W^+\right|})$ подсчитать количество его вхождений n_i в множество предложений S^+ и определить значение функции $n^+(w)$ в точке w_i так, что $n^+(w_i)=n_i$. Пусть $n^+(w)=0$ для словосочетаний $w \notin W^+$.

Шаг 4. Выполнить синтактико-семантический разбор каждого предложения множества S^-, расширяя множество W^- словосочетаниями с синтаксическими и семантическими связями.

Шаг 5. Для каждого встретившегося словосочетания w_j $(j=\overline{1,\left|W^-\right|})$ подсчитать количество его вхождений m_j в множество предложений S^- и определить значение функции $n^-(w)$ в точке w_j так, что $n^-(w_j)=m_j$. Пусть $n^-(w)=0$ для словосочетаний $w \notin W^-$.

Шаг 6. Сформировать множество W путем добавления в него словосочетаний $w \in W^+$, для которых выполняются неравенства $n^+(w) > n^-(w)$ и $n^+(w) > 1$.

Сложность алгоритма равна $O(|S^+|+|S^-|)$. Действительно, чтобы распределить все словосочетания по множествам W^+ и W^-, подсчитав количество вхождений каждого словосочетания, необходимо выполнить не более $n_{max} \cdot (|S^+|+|S^-|)$ операций, где n_{max} – наибольшее число словосочетаний, выделенных в пределах одного предложения. Для формирования множества W потребуется рассмотреть $|W^+|$ словосочетаний (при этом $|W^+| \le n_{max} \cdot |S^+|$), для каждого из которых выполняется фиксированное число действий при проверке неравенств. Таким образом, число операций пропорционально числу входных предложений, что говорит о линейной сложности алгоритма.

Блок-схема алгоритма формирования общенаучного словаря устойчивых словосочетаний представлена на рис. 6.

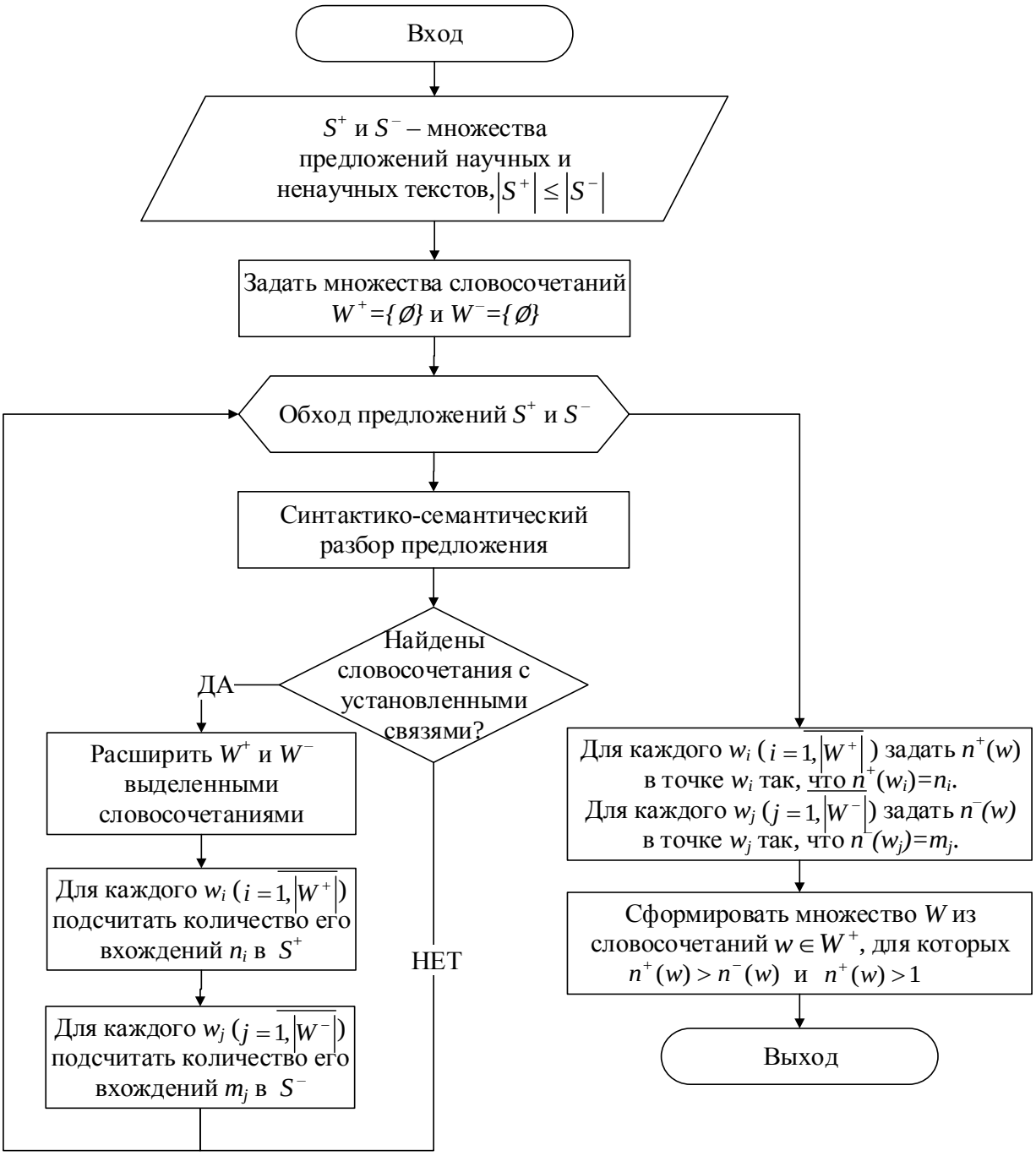


Рисунок 6 – Блок-схема алгоритма формирования общенаучного словаря устойчивых словосочетаний

В настоящей работе для формирования словаря использован Национальный корпус русского языка (НКРЯ) [57], который предназначен для обеспечения научных исследований лексики и грамматики языка. Корпус является представительным и содержит многие типы письменных и устных текстов, представленные в языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и

т.п.). В системе НКРЯ реализован лексико-грамматический поиск, с помощью которого можно находить предложения, содержащие слова запроса в произвольной словоформе.

С использованием существующих словарей общенаучной лексики вручную составлен список, содержащий более пятисот слов, многие из которых без контекста могут быть одновременно отнесены к различным стилям речи. С помощью средств системы НКРЯ составлены подкорпусы научной и ненаучной литературы, внутри которых выполнялся поиск слов полученного списка. Получено более миллиона предложений научного подкорпуса и более двух миллионов предложений ненаучного подкорпуса. Затем выполнялись шаги 1-6 представленного выше метода. В результате выявлены всевозможные словосочетания с синтаксическими и семантическими связями, которые входят в научный подкорпус чаще, чем в ненаучный, и встречаются более одного раза. Получено свыше 500 тысяч словарных единиц.

В табл. 1 представлены примеры словосочетаний полученного словаря – *w*. Второй столбец показывает число вхождений в подкорпус научной литературы НКРЯ – $n^{+}(w)$, третий столбец – отношение количества вхождений словосочетания в научный подкорпус к общему числу вхождений в научный и ненаучный подкорпусы – $P = \frac{n^{+}(w)}{n^{+}(w) + n^{-}(w)}$. В соответствии с алгоритмом 2.1 значения третьего столбца всегда больше 0,5.

Таблица 1. Примеры единиц сформированного словаря общенаучных словосочетаний

Словосочетание w	$n^{+}(w)$	P
принятие решений(я)	293	0,6
мировая практика	291	0,79
результат измерений(я)	192	0,89
объяснить явление	138	0,75
энергия активации	70	1,0
исследовать образец	39	0,91
заметно активизироваться	13	0,59
вызывать активацию	10	0,83
механизм активации	9	1,0
индуцировать активацию	7	1,0

На рис. 7 показана зависимость объема словаря от размера научного подкорпуса предложений. Значения горизонтальной оси соответствуют количеству проанализированных предложений научного подкорпуса. Значения вертикальной оси указывают количество словосочетаний, добавленных в словарь.

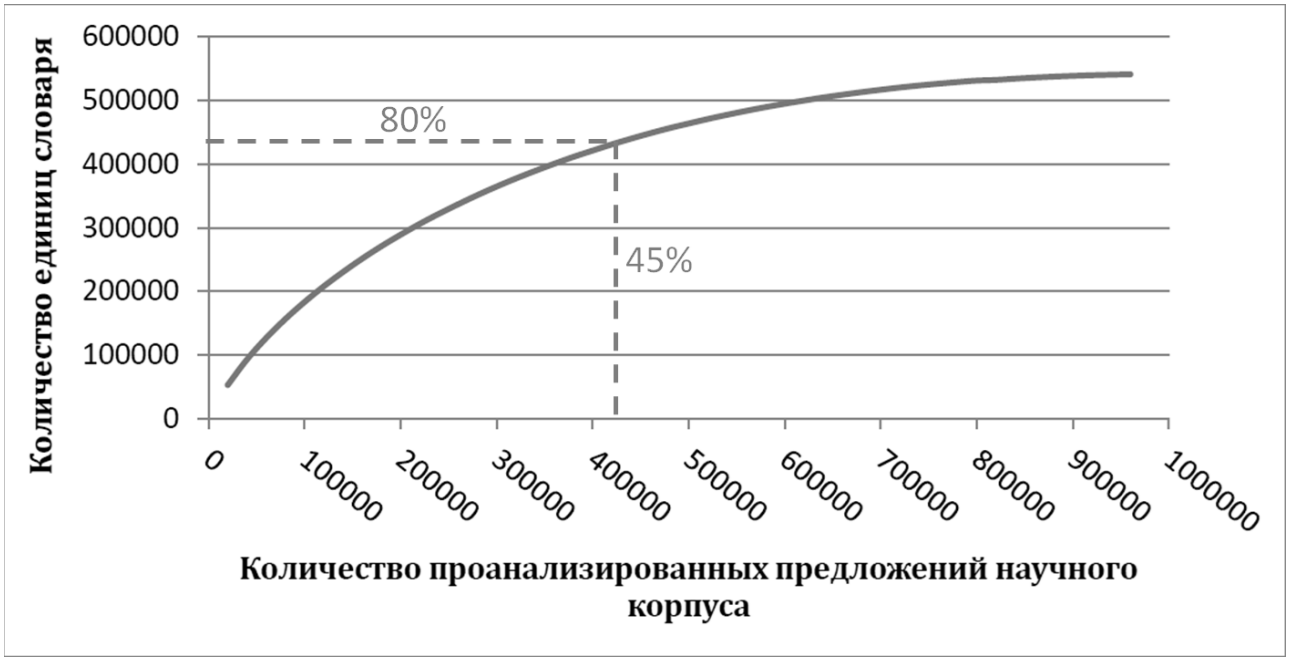


Рисунок 7 – Наполнение общенаучного словаря устойчивых словосочетаний

Из рис. 7 видно, что 80% словаря формируется при анализе 45% предложений подкорпуса, разбор каждого последующего предложения добавляет незначительное число словосочетаний. После 700 тыс. предложений скорость наполнения уменьшается, касательная к графику близка к горизонтальной. Это говорит о высокой полноте полученного словаря и о том, что одного миллиона предложений достаточно для того, чтобы покрыть наиболее часто встречающиеся в научной речи словосочетания.

На рис. 8 показано, какое количество различных единиц словаря встретилось в предложениях научного (белый цвет) и ненаучного (серый цвет) подкорпусов. Каждые 20 тысяч предложений научного подкорпуса содержат около 76 тысяч единиц словаря (53 тысячи различных), тогда как в таком же количестве предложений ненаучного подкорпуса встретилось лишь 23 тысячи единиц словаря (16 тысяч различных).

Рисунок 8 – Встречаемость единиц общенаучного словаря в подкорпусах

Рассмотрим алгоритм определения в тексте количества устойчивых общенаучных словосочетаний и установим численные значения низкого, заниженного и приемлемого количества словосочетаний в научном тексте, входящих в сформированный словарь.

2.1.4. Анализ встречаемости единиц словаря в текстах научной сферы

Определение доли словосочетаний текста *T*, являющихся устойчивыми общенаучными словосочетаниями множества *W*, выполняется по следующему алгоритму.

Алгоритм 2.2 **(алгоритм определения в тексте количества устойчивых словосочетаний общенаучного словаря).**

Шаг 1. Выполнить синтактико-семантический разбор каждого предложения текста *T*, и заполнить множество W_T словосочетаниями с синтаксическими и семантическими связями.

Шаг 2. Определить n_w – количество вхождений в текст словосочетаний $w \in (W_T \cap W)$.

Шаг 3. Определить N – количество всех словосочетаний, входящих в текст ($w \in W_T$).

Шаг 4. Вычислить отношение $\frac{n_w}{N}$ – доля словосочетаний текста *T*, входящих в словарь *W*, которая и характеризует количество устойчивых общенаучных словосочетаний.

Сложность алгоритма равна $O(N)$, поскольку требуется последовательно проанализировать *N* словосочетаний, входящих в текст *T*.

Блок-схема алгоритма определения в тексте количества устойчивых словосочетаний общенаучного словаря представлена на рис. 9.

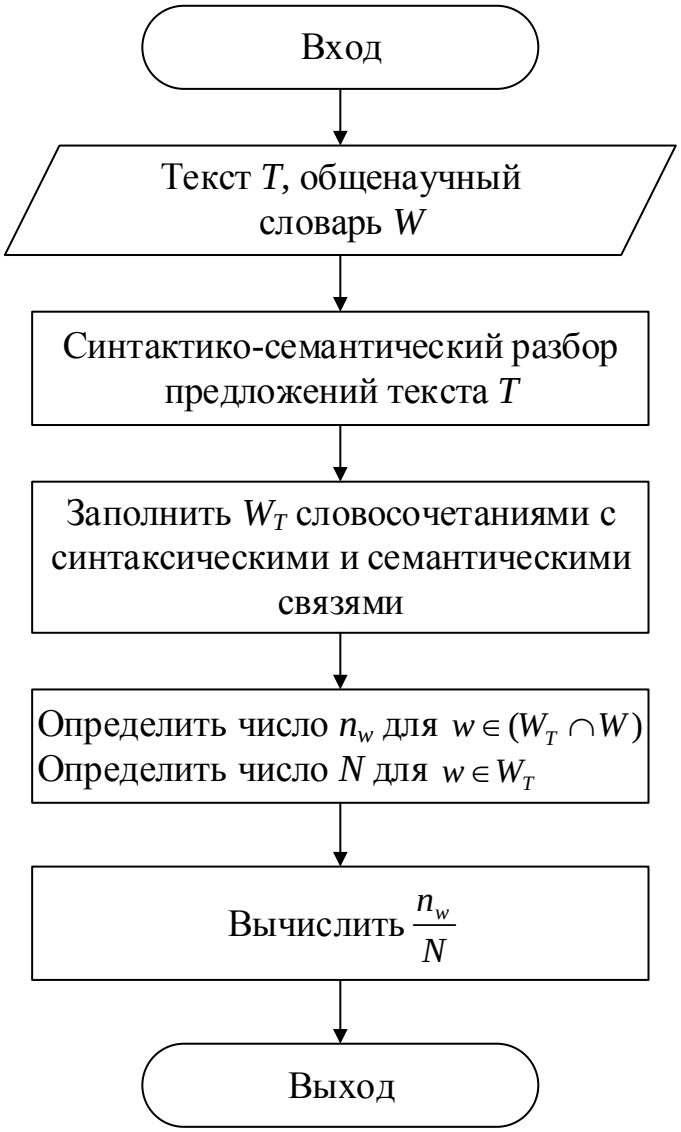


Рисунок 9 – Блок-схема алгоритма определения в тексте количества устойчивых словосочетаний общенаучного словаря

На корпусе научных текстов (более 40 тысяч статей), относящихся к различным научным направлениям, проведен анализ встречаемости полученных устойчивых словосочетаний. На рис. 10 для каждого текста представлено, какую его часть занимают словосочетания, которые входят в сформированный общенаучный словарь.

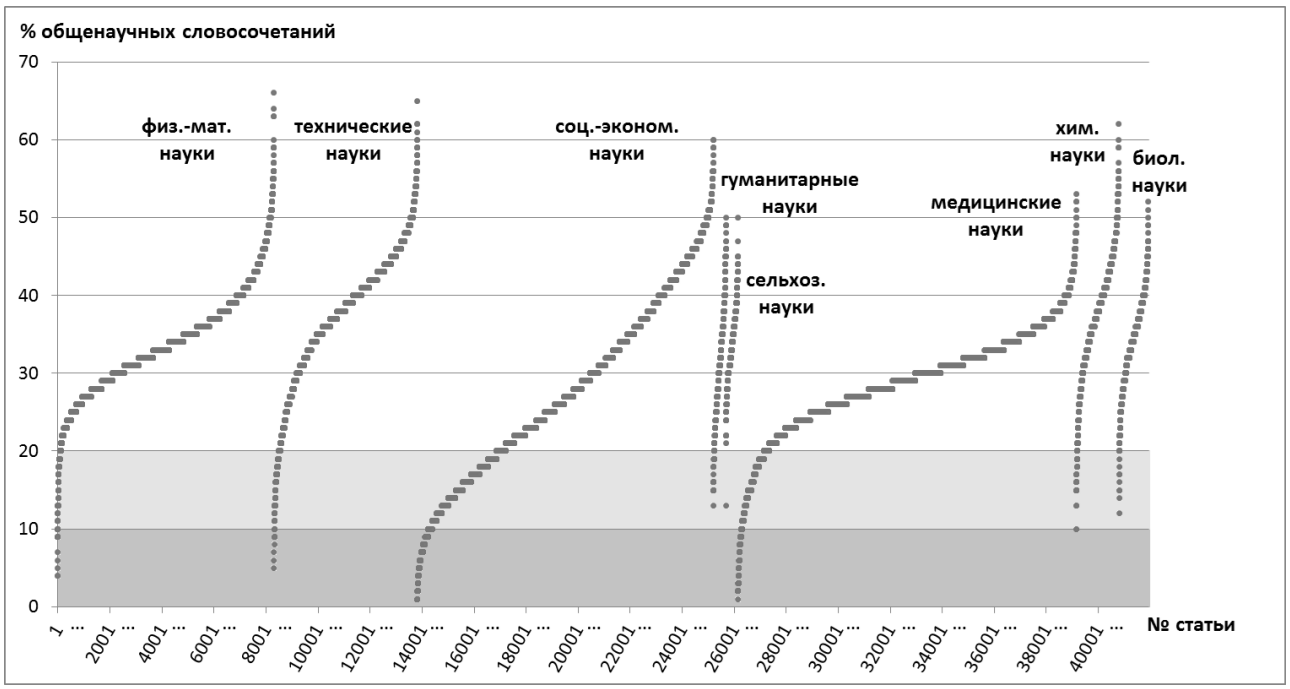


Рисунок 10 – Количество устойчивых общенаучных словосочетаний в текстах различных научных направлений

Примечание – статьи на графике (рис. 10) сгруппированы по научным направлениям и упорядочены внутри каждого направления по количеству устойчивых словосочетаний общенаучного словаря (ось ординат), вычисленному согласно алгоритму 2.2.

Полученные данные позволяют сделать вывод, что некоторые тексты социально-экономических и медицинских наук могут содержать небольшое число общенаучных словосочетаний – менее 20% среди всех словосочетаний текста, тогда как тексты остальных наук содержат от 20% и выше словарных единиц (за исключением незначительного числа публикаций). В связи с этим выделено три зоны (на рис. 10 обозначены разными оттенками): низкое количество общенаучных словосочетаний (0-10%), заниженное количество (10-20%) и приемлемое количество (20-100%). Эти значения позволяют отличить научные тексты от текстов, не относящихся к науке. Предлагается использовать установленные интервальные значения в качестве значений признака «количество устойчивых общенаучных словосочетаний», характеризующего качество научного текста.

Выполнено исследование применимости предложенного признака для выявления статей, не являющихся научными. Собраны коллекции ненаучных текстов (статьи журнала LiveJournal) и коллекция научно-популярных текстов (статьи журнала National Geographic). Выполнено автоматическое выявление общенаучных словосочетаний и вычислена доля, которую они занимают среди словосочетаний в каждом тексте. Установлено, что для большинства ненаучных текстов признак имеет значение «низкое количество» (рис. 11).

Рисунок 11 – Количество общенаучных словосочетаний в ненаучных текстах

Примечание – статьи на графике (рис. 11) упорядочены по относительному количеству общенаучных словосочетаний в тексте (ось ординат).

На рис. 12 представлены результаты эксперимента, показывающие, что научно-популярные статьи (наиболее близкие по стилю к научным статьям) также могут быть отделены от научных текстов с использованием полученного признака. Из графика следует, что большая часть статей имеет заниженное или низкое количество общенаучных словосочетаний.

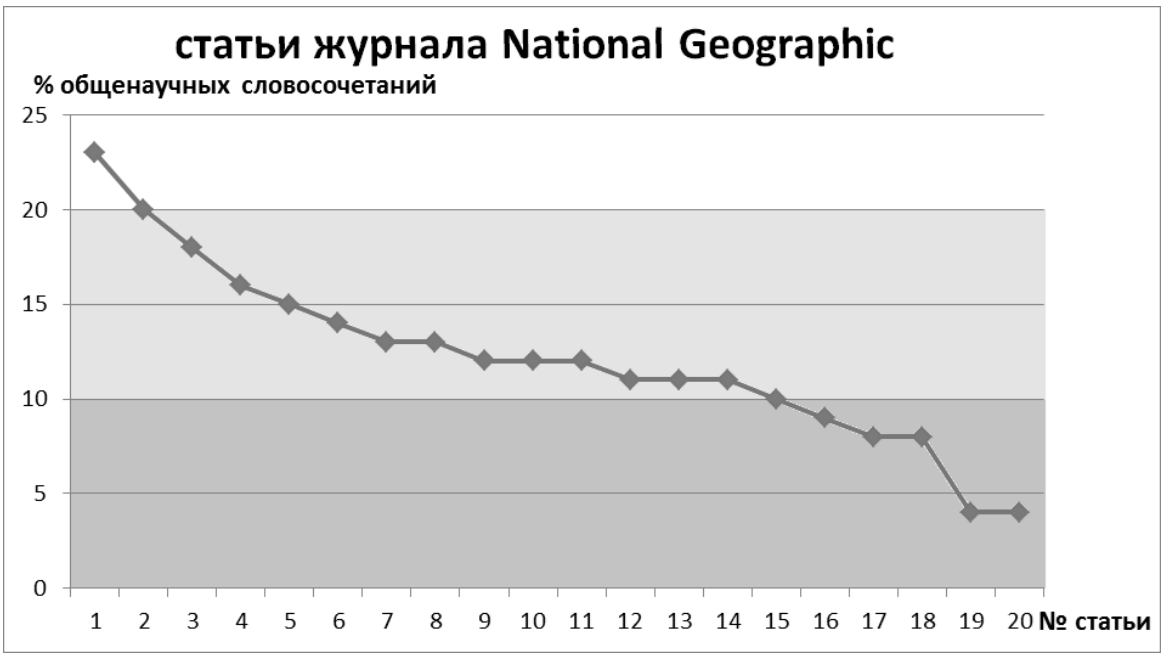


Рисунок 12 – Количество общенаучных словосочетаний в научно-популярных статьях

Примечание – статьи на графике (рис. 12) упорядочены по относительному количеству общенаучных словосочетаний в тексте (ось ординат).

На основе предложений научного и ненаучного подкорпусов НКРЯ сформировано множество возможных сочетаний слов русского языка, и, аналогично рассмотренному выше способу, определены интервалы приемлемого, завышенного и высокого количества специфических, необщеупотребимых словосочетаний.

Полученных признаков, однако, недостаточно, чтобы оценить качество текста. Существуют тексты с большим количеством общенаучных словосочетаний, но с неправильной структурой или большим числом лингвистических ошибок, поэтому требуется выделение дополнительных признаков. Рассмотрим далее метод определения структуры публикации.

2.2. Выявление структурных разделов в научной публикации

2.2.1. Выделение разделов формата IMRAD

В первой главе настоящей работы рассмотрена типичная структура научной публикации. Она включает в себя разделы, соответствующие формату IMRAD: «Постановка проблемы», «Методы», «Результаты», «Выводы».

Рассмотрим подробно методы, представленные в обзорной части настоящей работы, предназначенные для выявления перечисленных разделов.

В [39] предлагается классифицировать отдельные предложения. В качестве текстов выбраны биологические статьи, которые написаны в формате IMRAD и в явном виде содержат соответствующие формату разделы. Авторы вручную разметили коллекцию из 148 текстов и показали, что в действительности не все предложения соответствуют разделу, в котором они расположены, однако базовый классификатор, разделяющий предложения лишь по этому признаку, дает достаточно высокую точность 78,1%. В работе приведено сравнение мультиномиального наивного байесовского классификатора [40] и SVM-классификатора [37] и показано превосходство первого. В качестве признаков используются отдельные слова, двуграммы, триграммы и различные комбинации этих признаков. Показано, что наибольшая точность достигается при использовании всех трех типов признаков одновременно. Авторы предложили использовать дополнительные признаки, такие как времена глаголов, название раздела, в котором расположено предложение, маркеры цитирования, ссылки на таблицы и рисунки, числа, наличие отдельных стоп-слов. Первый признак выбран из предположения, что во введении присутствует больше глаголов настоящего времени, а в результатах – прошедшего. Экспериментально показано, что все признаки позволяют увеличить точность классификации до 91,95%. Наибольший прирост точности происходит, по мнению авторов, благодаря использованию информации о разделе, в котором находится предложение. Описанный метод хорошо работает на многих статьях биологических журналов, однако, далеко не во всех предметных областях принято выделять структурные элементы формата IMRAD в отдельные разделы в явном виде. На них базовый классификатор уже не будет давать такую высокую точность. К тому же, в некачественных статьях содержание текста может не соответствовать названию раздела, в котором он расположен, поэтому

использование этого признака, являющегося главным в описанном методе, некорректно для определения наличия структурных разделов.

В работе [41] также выполняется классификация отдельных предложений. Выбор предложения в качестве единицы классификации связан со спецификой задачи: результаты предполагается использовать для автоматического аннотирования статьи, которое принято выполнять путем объединения предложений. Как и в первой работе, корпус размеченных текстов был относительно небольшой и составлял 265 статей. Темы статей ограничивались двумя предметными областями – биохимией и химией. В качестве основных методов были выбраны метод опорных векторов с линейным ядром [37] и метод условных случайных полей [42]. Структура IMRAD была расширена и классификация выполнялась по 11 категориям: «Гипотеза», «Мотивация», «Цель», «Объект», «Фон», «Метод», «Эксперимент», «Модель», «Наблюдение», «Результат», «Вывод». Для классификации выбирались следующие признаки: положение предложения в статье (текст делится на некоторое число неравных по длине частей), положение предложения в разделе (раздел также особым образом делится на несколько неравных частей), заголовок раздела, положение в абзаце (абзац делится на 5 равных частей), длина предложения (соответствие одному из 9 заданных интервалов), число цитирований в предложении (возможны три случая – ни одного, одно и более одного), история (класс предыдущего предложения), униграммы, двуграммы и триграммы (при этом, как и в рассмотренной выше работе, стоп-слова оставлены), свойства глаголов (время, совершенный-несовершенный вид, личная-безличная форма), класс глагола (все глаголы предварительно распределяются по 10 классам), грамматические тройки <тип зависимости, главное, зависимое> (типы зависимости – субъект, прямой объект, косвенный объект и второй объект переходного глагола), наличие пассивного залога. В ходе экспериментов анализ признаков показал, что наибольшую роль играют n-граммы (особенно двуграммы), грамматические тройки, глаголы и такие признаки, как история классификации и заголовки

разделов. Получены не очень высокие значения F_1-меры. Самые высокие значения – 76% для категории «Эксперименты», 62% для категории «Фон», 53% для категории «Модель» и 51% для категории «Наблюдение». Для остальных категорий F_1-мера ниже 50%, самая низкая точность классификации получена для класса «Мотивация», F_1-мера составила 18%. В этом методе, как и в предыдущем, используются признаки, которые неприменимы для решения задачи определения наличия структурных разделов в статье. Однако некоторые признаки, например, такие как грамматические тройки, свойства глаголов, демонстрируют эффективность использования различных лингвистических характеристик текста, которые и предлагается учитывать при разработке метода в настоящей работе.

Опишем предлагаемый в настоящей работе метод выявления структуры научного текста. Некоторые результаты, связанные с разработкой метода, представлены в [4, 9]. Семантические и синтаксические конструкции, рассмотренные в разделе 2.1.2, могут быть использованы в качестве маркеров, описывающих способы оформления типичных структурных разделов первичного научного текста. Семантические конструкции, как было представлено, содержат предикатное слово в нормальной форме (чаще всего его замещает глагол) вместе со связанной с ним некоторой семантической связью синтаксической конструкцией (синтаксемой), замещающей определенную роль.

Для выявления маркеров используется корпус научных текстов, включающий статьи по биологии, физике, психологии, химии и медицине, размеченный экспертами с учеными степенями кандидатов наук: в каждом тексте вручную отмечались границы структурных разделов, не обязательно совпадающие с обозначенными в статье заголовками разделов. Посредством синтаксического и семантического анализа автоматически выделяются всевозможные словосочетания, при этом для каждого словосочетания подсчитывается число его вхождений в разные разделы. Каждое словосочетание становится маркером раздела с некоторой оценкой

принадлежности. Способ ее вычисления приведен ниже. Всего получено свыше 7 тысяч маркеров, относящихся к различным разделам публикации.

Примеры полученных семантических маркеров показаны в табл. 2, синтаксические маркеры (устойчивые словосочетания) представлены в табл. 3. В ячейках, отмеченных специальным символом «*», может находиться любое предикатное слово или синтаксема (в соответствии с названием столбца), столбец «*N*» содержит число вхождений данной конструкции в рассматриваемый структурный раздел, в следующем столбце «*P*» указан процент вхождений данного маркера относительно всех вхождений в различные части текстов, и в последнем столбце «*V*» указана оценка принадлежности маркера разделу.

Таблица 2. Примеры семантических маркеров структурного раздела «Постановка проблемы»

Предикатное слово	Синтаксема	Роль	N	P (%)	V
Являться	исследования	субъект	21	78	0,76
*	исследователей	агенс	15	88	0,85
Являться	работы	субъект	15	83	0,8
изучение	особенностей	объект	13	62	0,6
явиться	изучение	субъект	10	91	0,85
основываться	на *	источник_кауз_отн	9	69	0,66
являться	время	субъект	7	88	0,81
разрабатывать	*	объект	7	78	0,73
иметься	данные	посессив	7	64	0,6
иметься	в литературе	локатив	7	100	0,91
*	авторами	агенс	5	71	0,66
следовать	задачи	субъект	5	71	0,66
привлекать	внимание	субъект	4	100	0,85

Таблица 3. Примеры синтаксических маркеров структурного раздела «Постановка проблемы»

Устойчивое словосочетание	N	P (%)	V
цель работы	39	98	0,96
последний год	31	97	0,95
последнее время	19	90	0,88
в последнее десятилетие	9	100	0,93
такое исследование	6	100	0,89
ряд авторов	5	71	0,66
длительное время	5	63	0,58
широкое применение	4	100	0,85
число работ	4	100	0,85

Оценка принадлежности *V* вычисляется с помощью данных колонок «*N*» и «*P*» в таблицах и показывает с какой вероятностью появление того или иного маркера в новом анализируемом тексте соответствует наличию структурного раздела, описываемого этим маркером. Такие вероятности вычисляются с применением метода сглаживания Лапласа (Laplace smoothing или Additive smoothing) [58]. Это позволяет учесть неполноту исходных данных и изменить значимость маркеров с близкими значениями доли вхождений (колонка «*P*») так, что с большей вероятностью будет характеризовать раздел тот маркер, который встретился большее число раз. Так маркеры «цель работы» и «широкое применение» без сглаживания характеризуют раздел с вероятностями равными 0,98 и 1 соответственно (см. табл. 3). Применив метод сглаживания Лапласа с коэффициентом сглаживания 0,25, получим, что оценки принадлежности разделу маркеров равны 0,96 и 0,85, что говорит о том, что маркер «цель работы» лучше характеризует раздел «Постановка проблемы», чем маркер «широкое применение», хотя изначально вероятность при нем была меньше.

Если при анализе текста выявляется отсутствие всех маркеров одного из разделов, то это говорит об отсутствии в статье данного раздела. Наличие лишь маркеров, имеющих небольшую оценку принадлежности данному разделу,

также может говорить о его отсутствии. Исходя из этого, оценка соответствия некоторого текста структурному разделу вычисляется по совокупности маркеров этого раздела, входящих в текст, и зависит от маркера с максимальной оценкой принадлежности, от средней оценки принадлежности маркеров в тексте и от относительного количества встретившихся маркеров, соответствующих разделу:

$$E = \begin{cases} \mathrm{V}_{\max} - \dfrac{\mathrm{V}_{\max} - \mathrm{V}_{\mathrm{avg}}}{5}, \textit{если } \dfrac{\mathrm{n}}{N} > C \\ 0, \textit{иначе} \end{cases} \quad (1)$$

где n – число маркеров раздела в тексте с повторениями, N – общее число семантических и синтаксических конструкций в тексте, V_{max} – максимальная оценка принадлежности, V_{avg} – средняя оценка принадлежности, C – константа, задающая приемлемое относительное количество маркеров. Константа C задается на этапе обучения, своя для каждого структурного раздела. Формула (1) получена эмпирически, экспериментальная проверка формулы показала приемлемые результаты.

Опишем в общем виде алгоритмы выявления маркеров и определения наличия разделов. Положим, $M_I=\{\emptyset\}$, $M_M=\{\emptyset\}$, $M_R=\{\emptyset\}$, $M_D=\{\emptyset\}$ – множества, которые необходимо заполнить маркерами, характеризующими разделы «Постановка проблемы», «Методы», «Результаты» и «Выводы» (IMRAD) соответственно. Пусть S_I, S_M, S_R, S_D – множества предложений обучающей выборки, соответствующих указанным структурным разделам. Для выявления маркеров предлагается следующий алгоритм.

Алгоритм 2.3 **(алгоритм выявления маркеров структурных разделов).**

Шаг 1. Выполнить синтактико-семантический разбор каждого предложения множества S_I, расширяя множество M_I словосочетаниями с синтаксическими и семантическими связями.

Шаг 2. Повторить шаг 1 для множеств S_M, S_R, и S_D.

Шаг 3. Определить степень принадлежности разделу каждого маркера m_i множества M_I ($i=\overline{1,|M_I|}$), используя метод сглаживания Лапласа, $V_{m_i}=\frac{n_{m_i}+\alpha}{N_{m_i}+k\alpha}$, где n_{m_i} – число вхождений маркера m_i в множество предложений S_I, N_{m_i} – общее число вхождений маркера m_i во все предложения обучающей выборки, k – число различных разделов (в этом случае $k=4$), α – произвольный коэффициент сглаживания (положим, $\alpha=0{,}25$).

Шаг 4. Удалить из множества M_I маркеры со степенью принадлежности, не превышающей значение 0,5.

Шаг 5. Повторить шаги 3-4 для маркеров множеств M_M, M_R, и M_D. Маркеры построены.

Сложность алгоритма равна $O(|S_I|+|S_M|+|S_R|+|S_D|)$, поскольку число операций пропорционально числу входных предложений. На рис. 13 представлена блок-схема предложенного алгоритма.

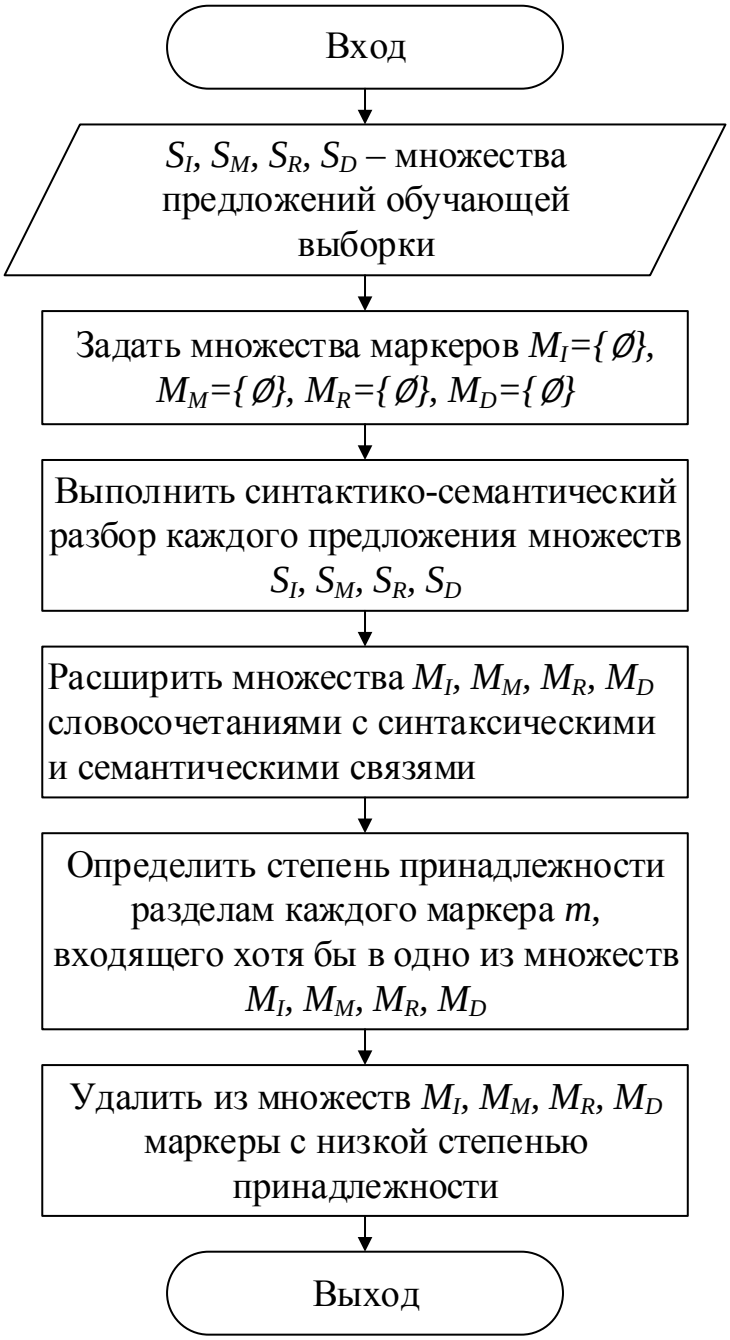


Рисунок 13 – Блок-схема алгоритма построения маркеров структурных разделов

Константа C в формуле (1) вычисляется один раз путем разделения обучающей выборки на две части, построения маркеров на текстах одной части и определения минимально возможного значения на текстах другой части.

Пусть T – произвольный текст. Для определения наличия в нем структурного раздела необходимо выполнить следующий алгоритм.

Алгоритм 2.4 **(алгоритм определения наличия структурного раздела в тексте).**

Шаг 1. Разделить текст T на фрагменты F_i равной длины, покрывающие весь текст.

Шаг 2. Выбрать один из фрагментов F. Для этого фрагмента выполнить синтактико-семантический разбор предложений и построить множество словосочетаний M_F.

Шаг 3. Найти пересечение множеств $M_F^I = M_F \cap M_I$.

Шаг 4. Вычислить значение E по формуле (1), используя степени принадлежности маркеров множества M_I, входящих в множество M_F^I.

Шаг 5. Повторить шаги 2-4 для каждого фрагмента F_i. Максимальное значение E и будем считать оценкой наличия раздела в тексте.

Сложность алгоритма равна $O(N)$, где N – число семантических и синтаксических конструкций в тексте. Действительно, для каждой конструкции необходимо последовательно выполнить фиксированное число операций, поэтому общее число операций пропорционально числу N. Блок-схема описанного алгоритма представлена на рис. 14.

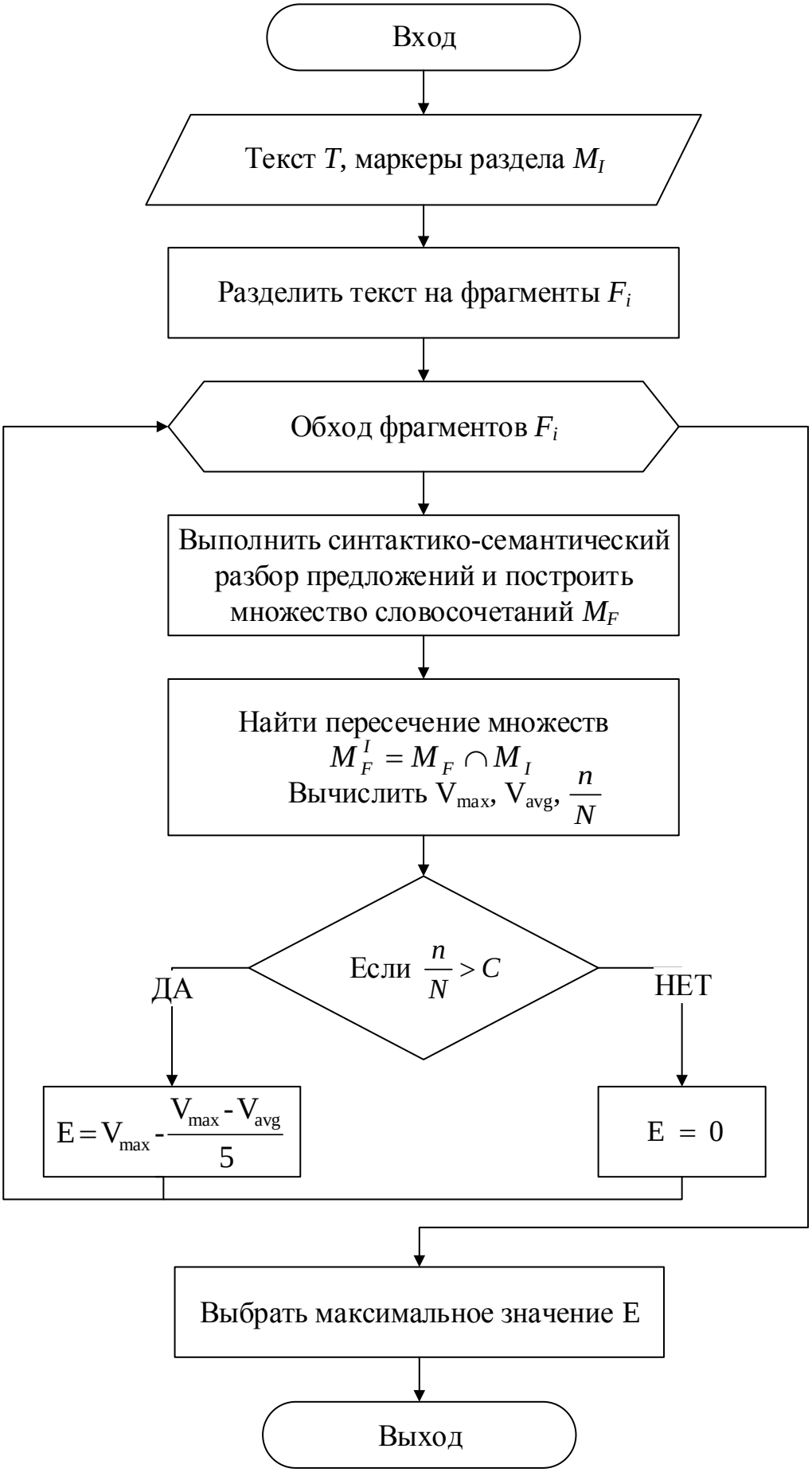


Рисунок 14 – Блок-схема алгоритма определения наличия структурного раздела

Экспериментально установлены три интервала значений переменной *E*, которые используются в качестве значений признака наличия раздела: [0; 0,5) – раздел, скорее всего, отсутствует, [0,5; p_i) – раздел, вероятно, отсутствует, [p_i; 1] – раздел присутствует. Значения p_i устанавливаются при обучении, отдельно для каждого типа структурного компонента. Эти значения признака используются в дальнейшем при индуктивном построении правил.

Проведено сравнение предложенного метода оценки наличия раздела с наивной байесовской классификацией на размеченной выборке из 200 статей с применением перекрестной проверки. Результаты сравнения представлены в табл. 4. Все значения качества классификации, такие как точность (precision), полнота (recall) и F_1-мера (F_1-measure) вычисляются согласно стандартным формулам [38].

Таблица 4. Результаты сравнения предлагаемого метода с наивным байесовским классификатором

	Наивная байесовская классификация			Предлагаемый в настоящей работе метод		
	Точность	Полнота	F_1-мера	Точность	Полнота	F_1-мера
«Постановка проблемы»	0,65	0,93	**0,76**	0,62	0,92	**0,74**
«Методы»	0,84	0,83	**0,83**	0,76	0,84	**0,8**
«Результаты»	0,6	0,9	**0,71**	0,63	0,93	**0,74**
«Выводы»	0,94	0,07	**0,13**	0,33	0,84	**0,47**

Установлено, что байесовский классификатор с высокой полнотой (0,8-0,9) определяет классы для разделов «Постановка проблемы», «Методы» и «Результаты», но с низкой полнотой определяет класс «Выводы», поэтому общая полнота классификатора не очень высокая и равна 0,68. Это связано с тем, что все разделы в обучающей выборке содержат схожие формулировки выводов, используемые при описании некоторых промежуточных выводов. Предложенный в настоящей работе метод позволяет оценивать наличие всех разделов в пределах одного анализируемого фрагмента текста, что значительно повышает полноту и значение F_1-меры для класса

«Выводы». Так если выводы будут приведены при описании результатов, то для этого текста будет определено и наличие результатов, и наличие выводов.

Рассмотрим метод, позволяющий выявлять признаки, характеризующие качество текстов научной сферы, связанные со свойствами списка использованных источников в публикации.

2.2.2. Выделение и структурирование списка литературы

Наличие списка использованной литературы является обязательным условием для принятия статьи к публикации, однако встречаются работы, у которых список литературы отсутствует, либо оформлен так, что не всегда возможно восстановить, на какие источники ссылается автор. Список источников также может косвенно говорить о знакомстве автора с темой его работы: если автор публикации вместо научной литературы использует неавторитетные источники (например, интернет сайты) или ссылается по большей части лишь на свои работы, то содержание публикации может быть необъективным. Кроме того, непринято необоснованно добавлять в список литературы источники, на которые нет ссылок внутри текста. Если в работе все источники являются необоснованно добавленными, то такая работа не может быть полезной читателю, поскольку становится невозможно определить, какой текст является авторским, что из описанного является новым, как пригодились автору указанные им источники.

Существующие методы структурирования списка использованных источников представлены в [59, 60]. В первой работе предлагается система извлечения цитируемых источников, которая для обнаружения и сегментации списка использованных источников применяет различные эвристики. Авторы второй работы предлагают выполнять извлечение источников, используя дополнительно информацию о макете статьи (размещение текста на печатной странице) и о форматировании – стилях и размерах шрифта. Недостатком обоих методов является возможность обрабатывать только правильно оформленные списки литературы: первым действием оба метода выполняют поиск заголовка,

предшествующего списку источников, при отсутствующем заголовке методы не работают.

Открытый программный продукт ParsCit [59] тестировался в рамках настоящей работы на российских и зарубежных публикациях с различными способами оформления списка литературы. В результате установлено, что имена авторов выделяются неправильно, в тех случаях, когда фамилии не являются общеизвестными, когда нарушена или отсутствует нумерация источников или выполняется анализ текста, представленного в виде одной строки без переносов. В связи с недостатками современных методов в рамках настоящей работы разработан метод, справляющийся с вышеуказанными проблемами.

Предложенный метод опирается на набор регулярных выражений, которые разработаны, протестированы и уточнены на большом числе примеров с разными способами представления списка источников. Регулярные выражения учитывают наиболее частые ошибки, допускаемые авторами при оформлении, а также возникающие при конвертации текстов из различных компьютерных форматов.

Метод заключается в автоматическом выявлении источников, структурировании каждого источника с извлечением имен авторов и редакторов, названия источника, даты опубликования, интернет-ссылки, и последующем анализе извлеченного списка литературы. Выполняется проверка соответствия количества уникальных ссылок, приведенных в содержательной части работы, количеству источников, указанных в списке литературы. Так выявляется наличие необоснованно добавленных источников, на которые нет ссылок внутри текста. Затем вычисляется отношение числа интернет сайтов к общему числу источников и доля источников, принадлежащих каждому автору. Если большая часть источников принадлежит одному автору, на это следует обратить внимание: как правило, это говорит о неоправданно высоком самоцитировании, которое может быть следствием низкой осведомленности автора о положении дел в его предметной области. Дополнительно в тексте

выявляются ссылки, указывающие одновременно на большое число источников, например “*[2-18]”*. Если автор не цитирует отдельно источники, указанные в такой ссылке, это может говорить о необоснованном наполнении списка литературы для придания значимости работе.

Итак, метод выделения структурных разделов и метод анализа списка использованных источников, предложенные в работе, позволяют выделить следующие признаки, характеризующие качество текста:

- наличие разделов, соответствующих формату IMRAD;
- наличие списка источников;
- авторитетность источников;
- процент самоцитирования;
- наличие необоснованно добавленных источников.

2.3. Обнаружение лингвистических ошибок в научных текстах

2.3.1. Описание метода обнаружения лингвистических ошибок

В рамках настоящей работы разработан метод обнаружения лингвистических ошибок в научных текстах. Некоторые полученные результаты, связанные с разработкой метода, изложены в [3]. В основе метода лежит некоторое множество правил *R*, с помощью которых можно выявить нарушения правил согласования, нарушения семантической связности, последовательности изложения и др. Для формирования множества правил предлагается следующий алгоритм.

Алгоритм 2.5 **(алгоритм формирования правила, характеризующего лингвистические ошибки).**

Шаг 1. Выбрать одно из правил русского языка *r′*.

Шаг 2. Исследовать примеры предложений из множества S^+, удовлетворяющих данному правилу, и примеры предложений с нарушением правила из множества S^-.

Шаг 3. Извлечь условия, выполнение которых свидетельствует о наличии ошибки. При формировании условий степень обобщения ограничивается множеством правильных предложений S^+.

Шаг 4. В выборке научных текстов выделить предложения S^0, для которых выполняются полученные условия.

Шаг 5. Если среди выделенных предложений содержатся правильные предложения ($S^0 \cap S^+ \neq \emptyset$) или обнаруживаются предложения с нарушениями S^-, которые не были выделены ($S^- \setminus S^0 \neq \emptyset$), и есть возможность уточнить условия так, чтобы правило покрывало меньше предложений из S^+ и больше из S^-, то уточнить правило и выполнить шаг 4. Правило r является результатом последовательного итерационного уточнения условий.

Блок-схема алгоритма формирования правила представлена на рис. 15.

Рисунок 15 – Блок-схема алгоритма формирования правила

Общий метод обнаружения лингвистических ошибок в научных текстах заключается в последовательном применении сформированных правил к результатам лингвистического анализа, которые представляют собой полуструктурированные данные, т.е. текст с установленными свойствами его элементов. Такие данные позволяют учитывать семантические, синтаксические, морфологические и лексические характеристики элементов текста, их контекст и взаимную сочетаемость. Лингвистический анализ проводится на первом шаге алгоритма обнаружения ошибок. В ходе работы алгоритма, при выполнении условия какого-либо правила, предложение, содержащее нарушение, добавляется в структурированный список подозрительных предложений вместе с меткой типа ошибки. Одновременно увеличивается показатель количества выявленных ошибок соответствующего типа. Такие показатели будут использоваться для определения значений признаков, характеризующих качество текста научной сферы. В настоящей работе признаки соответствуют различным нарушениям и имеют значения «нарушение присутствует», «нарушение отсутствует». Опишем алгоритм обнаружения лингвистических ошибок в общем виде.

Алгоритм 2.6 **(алгоритм обнаружения лингвистических ошибок).**

Шаг 1. Выполнить лингвистический анализ предложений текста *T*.

Шаг 2. Применить правила r_i ($i = \overline{1,|R|}$, где R множество правил) к результатам лингвистического анализа, последовательно для каждого предложения.

Шаг 3. Добавить предложения, для которых выполняются условия правил, в множество предложений S^-, и задать значение каждой переменной n_i, равной числу найденных нарушений по правилу r_i.

Шаг 4. Определить номинальные значения признаков, характеризующих качество текста, на основе количества нарушений n_i.

Сложность алгоритма равна $O(N \cdot |R|)$, где N – число предложений, поскольку необходимо рассмотреть каждое из предложений $|R|$ раз для проверки каждого правила. Блок-схема алгоритма представлена на рис. 16.

Рисунок 16 – Блок-схема алгоритма обнаружения лингвистических ошибок

Рассмотрим более подробно примеры лингвистических ошибок и полученные правила для их обнаружения.

2.3.2. Обнаружение нарушений правил согласования

Одной из распространенных ошибок, встречающихся в научных текстах, является отсутствие согласования причастия с определяемым словом, стоящим перед причастным оборотом. Приведем сформированное правило, позволяющее выявлять такие ошибки.

Правило 1: если в состав предложения входит причастный оборот, выделенный запятыми, и причастие не согласуется ни с одним из существительных, местоимений, прилагательных и числительных, стоящих перед оборотом, в роде, числе и падеже в ед. ч. и в числе и падеже во мн. ч. (и не согласуется с однородными членами в падеже во мн. ч.), то такое предложение содержит нарушение согласования.

Приведем пример предложения, удовлетворяющего данному правилу. «Существует возможность превращения идиолекта в некий субстрат с аморфным *содержанием* и «экономной» *формой*, *не дающих* реальных шансов для диагностики говорящего». Как видно, отсутствует падежное согласование причастия «дающих» с однородными определяемыми словами «содержанием» и «формой».

Близким нарушением является неоднозначность связи причастий с определяемыми словами в причастных оборотах. Соответствующее правило звучит следующим образом.

Правило 2: если в состав предложения входит причастный оборот, выделенный запятыми, и причастие не согласуется с примыкающим слева к причастному обороту существительным и, а) если число причастия единственное и число существительных, стоящих перед причастным оборотом, согласующихся с ним в роде и падеже, больше одного, или б) если число причастия множественное и б1) есть два подходящих существительных во множественном числе, не являющихся однородными членами, или б2) есть

пара существительных, являющихся однородными членами, и несвязанное с ними существительное во множественном числе, которые согласуются с причастием в падеже, то предложение содержит нарушение, состоящее в неоднозначности связи причастия с определяемыми словами.

Рассмотрим пример, удовлетворяющий правилу 2. «Причем, *масса* и *энергия* элементарной волны фотона, *определенная* из уравнения …, будет составлять соответственно…». Причастие «определенная» может относиться как к слову «масса», так и к слову «энергия». Кроме того, возможно, была неправильно выбрана форма причастия. Такая многозначность толкования и является следствием допущенного нарушения. Приведем еще один пример. «Соотношение *неопределенностей энергии* стержня и *времени*, *рассчитанных* по однокоординатным данным». В данном случае невозможно определить, к чему относится причастие «рассчитанных»: к сущ. во мн.ч. «неопределенностей» или к однородным членам, выраженным существительными «энергии» и «времени».

Существуют правила употребления подчинительной связи прилагательного. При этом различают предложное и беспредложное управление. Отдельное правило обычно задается управляемым падежом с предлогом или без него (в случае беспредложного управления) и перечислением конкретных прилагательных-лексем [61]. Например, к беспредложному управлению прилагательных относятся следующие связи:

– род. п.: *чуждый чего-н., достойный кого-чего-н., полный чего-н.*;

– дат. п.: *подобный, присущий кому-чему-н., пропорциональный, равносильный чему-н., чуждый, противный кому-н.*;

– тв. п.: *богатый, обильный, известный, характерный чем-н.*

К предложному управлению относится, например, управление следующими падежами:

– дат. п. с предлогом *к*: *склонный, способный, восприимчивый, чувствительный к чему-н., крайний, близкий к чему-н.*;

– вин. п. с предлогом *на*: *ловкий, проворный, смелый на что-н., похожий на кого-что-н.*;

– тв. п. с предлогом *с*: *единый, сходный, общий, одинаковый с кем-чем-н., близкий с кем-н., смежный, соседний, однородный, сопряженный с чем-н.*

Для выявления нарушения подчинительной связи прилагательного построено следующее правило.

Правило 3: если в предложение входит прилагательное, для которого есть правила управления, и следующее за ним существительное (с предлогом или без него) не соответствует ни одному из правил и не согласуется с прилагательным в роде, числе и падеже, то предложение содержит нарушение правила употребления подчинительной связи прилагательного.

Приведем примеры предложений, для которых выполняется условие правила. «Плюс, какие более мелкие полевые структуры могут в конкретных структурах микромира, обладая *соответствующими масс* – энергетическими параметрами, выполнять функции…». «Из (26) подставляем в (23), получим (27), где gm – заряд *эквивалентный диполя* эфира».

Частой ошибкой является нарушение согласования сказуемого с однородными подлежащими. Согласно [62] при прямом порядке слов (сказуемое следует за однородными подлежащими) обычно употребляется форма множественного числа сказуемого, при обратном порядке (сказуемое предшествует подлежащим) – форма единственного числа, однако в зависимости от формы связи между однородными подлежащими это правило может изменяться. Например, при смысловой близости однородных подлежащих предпочитается форма единственного числа сказуемого. Сказуемое однозначно ставится во множественном числе лишь в случае, когда возникает необходимость согласования в роде и подлежащие принадлежат к разному грамматическому роду. В связи с этим можно сформулировать следующее правило выявления предложений с ошибками.

Правило 4: если в состав предложения входят однородные подлежащие, принадлежащие к разному грамматическому роду, и сказуемое в форме глагола

прошедшего времени единственного числа, то предложение содержит нарушение согласования сказуемого с однородными подлежащими.

Приведем примеры предложений, удовлетворяющих данному правилу: «Именно такой философией *являлся* для своего времени *марксизм* и *философии*, лежащие в основе буржуазных революций». «Несмотря на то, что все преобразования … существовали в разных видах в разных местах, … *выбор* и *ответственность* за него *ложился* на реформатора». В каждом примере курсивом выделены однородные подлежащие и не согласованное с ними сказуемое.

При построении конструкций управления важен правильный выбор падежных форм. Например, предлоги «согласно», «вопреки» управляют дательным падежом (согласно, вопреки приказу) и творительным (согласно с требованиями), постановка существительного в родительном падеже недопустима [62]. При нарушении падежа в результате лингвистического анализа предлог «согласно» не будет связан с управляемым словом, и часть речи для слова «согласно» может быть установлена как наречие. Поэтому признаком ошибки является наличие несвязанного с другими элементами предложения слова «согласно» и следующего за ним слова в родительном падеже. Ниже представлено правило, с помощью которого можно выявлять такое нарушение согласования.

Правило 5: если предложение содержит лексический элемент «согласно» или «вопреки», несвязанный синтаксически с другими элементами предложения, и первое слово, которое расположено после этого элемента и непосредственно после следующих за ним наречий, союзов и числительных в числовой записи, имеет характеристику «падеж» и стоит в родительном падеже, то такое предложение содержит нарушение согласования.

Такая сложная структура правила необходима для того, чтобы не пропустить предложения, в которых слово с родительным падежом не является первым, например, «Согласно 4 прянятых законов…», и в то же время, чтобы

не отнести к подозрительным правильные предложения, например, «Согласно [5], нет необходимости выполнять…».

Еще одним признаком нарушения грамматической нормы является следование за словами «более» и «менее» сравнительной или превосходной степени прилагательного. Опишем правило для выявления такой ошибки.

Правило 6: если предложение содержит слово «более» или «менее» и следующее за ним слово является прилагательным в сравнительной или превосходной степени, то предложение содержит нарушение согласования.

Примерами нарушений являются, например, словосочетания «более энергичнее», «более оптимальный».

2.3.3. Обнаружение нарушений синтаксической и семантической связности

В разделе 1.1.4 показано, что одним из нарушений является большое число синтаксически или семантически несвязных слов. Такие слова могут быть обнаружены в результате синтаксического и семантического анализа: они отделяются от синтаксического дерева (отсутствует связь со словом-родителем) и не входят в семантическую сеть.

В связи с тем, что уровень современных лингвистических анализаторов не позволяет безошибочно строить синтаксическое дерево, не представляется возможным делать выводы о наличии нарушения для отдельного предложения. Однако на большом числе предложений может быть установлено допустимое число несвязных слов, превышение которого будет говорить о низкой синтаксической и семантической связности. На обучающей выборке научных статей установлен порог приемлемой доли слов в тексте, для которых лингвистическим анализатором не установлена связь со словом-родителем. Для обнаружения нарушения связности текста сформировано следующее правило.

Правило 7: если в тексте превышено допустимое количество слов, не связываемых со словами-родителями, то степень синтаксической и семантической связности текста является низкой.

2.3.4. Обнаружение лексической избыточности

Как отмечено в разделе 1.1.5, лексическая избыточность (неоправданное многословие) является нарушением норм лексической стилистики и образуется вследствие излишнего употребления плеоназмов. Плеоназмы, содержащиеся в тексте, можно выявить путем подсчета количества одинаковых слов, употребляемых в пределах одного предложения. Анализ научных текстов показал, что лексической избыточностью могут обладать предложения с плеоназмами, содержащими более двух слов с совпадающими нормальными формами, например, следующее предложение: «На *первых* этапах обучения на *первый* план выходят *первые* два аспекта». Иногда употребление таких плеоназмов оправдано, однако текст не должен быть перенасыщен ими. Проведена оценка уровня встречаемости плеоназмов в научном тексте, которая показывает, что количество слов, образующих плеоназмы, не превышает 0,75% от количества всех слов текста. При этом не учитываются союзы, предлоги, частицы, числа и слова, для которых не определяется часть речи (например, элементы математических формул). Приведем составленное правило для выявления этого нарушения.

Правило 8: если в тексте превышено допустимое число слов, образующих плеоназмы, не являющихся союзами, предлогами, частицами, числами или словами, для которых не определяется часть речи, то текст обладает лексической избыточностью.

2.3.5. Обнаружение нарушений последовательности изложения

В разделе 1.1.6 отмечалось, что нарушение последовательности изложения может характеризоваться отсутствием одного из парных элементов, предназначенных для обозначения порядка явлений и связей между ними. К ним относятся словосочетания «с одной стороны, с другой (стороны)», или вводные слова «во-первых, во-вторых». Опишем правило для выявления такого нарушения.

Правило 9: имеет место нарушение последовательности изложения в тексте, если выполнено, по крайней мере, одно из условий:

1) второй элемент пары встретился в тексте раньше, чем первый;

2) между однотипными элементами определенной пары отсутствует элемент другого типа, т.е. пропущен один из элементов;

3) после первого элемента пары в оставшейся части текста отсутствует второй элемент.

При отсутствии одного из элементов стоит обратить внимание: возможно, если автор не использовал других средств, заменяющих эти элементы, нарушен порядок изложения, например: «*Нашим восприятием знаков, напоминающих нам об истории, репрезентирующих те или иные события в актуальном настоящем, управляют несколько важных механизмов. Во-первых, это «распознавание имени»*....». Другие «важные механизмы» не отмечены явно в тексте, что затрудняет его целостное восприятие.

2.3.6. Результаты применения метода автоматического обнаружения лингвистических ошибок

Для проверки эффективности выявления предложений, содержащих нарушения правил русского языка, с помощью разработанного метода обнаружения лингвистических ошибок и алгоритма 2.6 проанализировано свыше 600 публикаций, которые в основном представляют собой статьи научных студенческих конференций. Результат выполнения алгоритма показал, что среди выделяемых по правилам предложений содержится высокий процент предложений с нарушениями. Примеры ошибок, выявленных автоматически в соответствии с предложенными правилами, приведены в табл. 5. Средняя колонка «*k*/*N*» содержит два показателя: N – общее число автоматически выявленных предложений, удовлетворяющих сформированным правилам, и k – количество предложений, действительно содержащих нарушения.

Таблица 5. Примеры автоматически выявленных лингвистических ошибок

Тип нарушения	**k/N**	**Примеры предложений с нарушениями**
1. Отсутствие согласования причастия с определяемым словом	15/98	С учетом ранее полученных данных, о том, что пассивные дети чаще встречаются в семьях недостаточно стимулирующего типа, можно предположить, что экспериментатор пытается компенсировать тип взаимодействия, ***сложившейся*** в семье, побуждая ребенка к активным действиям.
		Была составлена анкета для опроса жителей, которая включала спектр вопросов по отраслям (ЖКХ, потребительский рынок, образование, здравоохранение, культура, спорт, социальная поддержка населения и т. д.), ***оценивающую*** инфраструктуру данного города.
2. Нарушение последовательности вводных слов «во-первых, во-вторых»	15/15	…перед испытуемыми стоит задача, ***во-первых***, методом проб и ошибок найти значимые клавиши, удержать их в памяти, ***а затем***, согласно инструкции, осветить ячейки в порядке возрастания цифр.
		Стремление к общению приглушено по двум причинам. ***Во-первых***, высокая критичность к другим не способствует накоплению позитивного опыта общения… (*вторая причина не указана*).
3. Нарушение последовательности «с одной стороны, с другой (стороны)»	40/58	Отсутствие реального опыта собственного материнства, осмысление опыта родительской семьи ориентирует девушек, ***с одной стороны***, на формирование с будущим ребенком доверительных, близких, товарищеских отношений. (*продолжение мысли отсутствует*)
		В данном контексте не следует рассматривать Европейский союз как иерархичную систему управления. ***С одной стороны***, сама логика возникновения... ***Однако*** Европейский союз, в отличие от национального государства, — это скорее переговорная система...

Тип нарушения	k/N	Примеры предложений с нарушениями
4. Неправильный выбор падежной формы после предлогов «согласно», «вопреки»	9/12	Все работы по расконсервации скважины проводятся ***согласно типовых правил*** и ***инструкций***, с противовыбросовым оборудованием и герметизирующей головкой, установленными на скважине для предотвращения аварийного выброса нефти.
		Урожайность кукурузы в большей степени, ***согласно наших расчетов***, зависит от осадков за май месяц (r > 0,5), потом идет сумма осадков за май и июнь месяцы.
5. Сравнительная степень прилагательного после слов	2/4	Динамика разрушения ПДС и восстановления подвижности воды происходят в карбонатах ***более медленнее,*** чем в кварцевых моделях.
		Более ярче это проявилось в карбонатных пористых средах.

Поясним данные второго столбца в первой строке таблицы. При проверке выполнения условия первого правила было автоматически проанализировано свыше 12 тыс. предложений, содержащих причастный оборот. Из этого множества предложений было отобрано всего 98 предложений, удовлетворяющих условию правила, что составляет около 1% от всех предложений с причастными оборотами, встречающихся в текстах данной выборки. Таким образом, для выявления предложений с ошибками такого типа требуется рассмотреть лишь 98 предложений, оставшиеся же 12 тыс., благодаря методу, исключаются автоматически, что значительно упрощает работу по выявлению таких нарушений.

Что касается остальных полученных числовых данных – видно, что точность автоматического обнаружения нарушений достаточно высокая. Так, условие правила выявления нарушений последовательности вводных слов «во-первых, во-вторых» выполнялось только в случае наличия ошибки (15/15). В отношении нарушений при выборе падежной формы существительного после

предлогов «согласно» и «вопреки» также достигается достаточно высокая точность (9/12).

Предложенный метод позволяет выявлять следующие признаки, характеризующие качество текста:

- нарушения согласования прилагательных с существительными в роде, числе и падеже;
- нарушения подчинительной связи прилагательного;
- нарушения согласования сказуемого с однородными подлежащими;
- нарушения согласования причастий с определяемыми словами в причастных оборотах;
- неоднозначность связи причастий с определяемыми словами в причастных оборотах;
- употребление превосходной степени прилагательного;
- количество нарушений синтаксической и семантической связности;
- лексическая избыточность;
- нарушения последовательности изложения.

Результаты главы 2

В настоящей главе предложен комплекс методов для выявления признаков, характеризующих качество текста научной сферы, а именно:

- метод построения общенаучного словаря устойчивых словосочетаний;
- метод автоматического выявления структурных разделов научной публикации;
- метод автоматического обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.

Все методы используют лексические, морфологические, синтаксические, семантические и различные информационные характеристики текста. Такие характеристики позволяют оперировать с текстом не как с бессвязным набором слов, но как с полуструктурированными данными, учитывая связи и отношения между элементами текста.

Разработанные методы реализованы в виде программных модулей анализа качества текста и внедрены в программный комплекс интеллектуального поиска и анализа научных публикаций «Exactus Expert» [63], с помощью которого выполнялось тестирование предложенных методов. Описание программных модулей, а также снимки системы «Exactus Expert» и примеры отчетов, получаемых в результате работы программ, приведены в Приложении 1.

Проведенные эксперименты подтверждают, что разработанные методы применимы для обнаружения различных нарушений и отступлений от норм научного текста. Покажем в следующей главе, что полученное множество признаков может быть использовано для решения задачи выявления псевдонаучных текстов.

Глава 3. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫЯВЛЕНИЯ ПРИЗНАКОВ ПСЕВДОНАУЧНЫХ ТЕКСТОВ

В рамках настоящей работы решалась задача выявления псевдонаучных текстов с целью проверки применимости предложенных в главе 2 признаков для определения качества текстов научной сферы.

Обнаружение псевдонаучных текстов предлагается проводить путем решения задачи классификации, имеющей следующую формулировку. Пусть множество $O=\{o_i\}$ – множество объектов, объектом в данном случае является текст научной сферы; множество $P=\{p_j\}$ – множество признаков, каждый из которых обладает своим множеством допустимых значений $p_j^1,...,p_j^{n_j}$. Один объект обладает одним значением каждого признака, которое называется свойством объекта. Каждый объект представляется в виде вектора свойств $o = \{ p_1^{i_1},...,p_n^{i_n} \}$. Необходимо найти отображение $F:O\to K$, где $K=\{k_1,k_2\}$ – классы объектов, один из которых соответствует псевдонаучным, а другой – научным текстам. Такое отображение может быть построено с помощью методов машинного обучения на некотором множестве публикаций.

В настоящей главе большинство представленных экспериментов проводилось на текстах сформированной коллекции, состоящей из 2131 псевдонаучной публикации и 938 научных статей рецензируемых журналов, входящих в перечень ВАК. Научные публикации относились к техническим и медицинским наукам, которые были выбраны как наиболее близкие к темам псевдонаучных работ.

Следующая часть работы посвящена решению ряда задач: разработке метода обнаружения псевдонаучных фрагментов, формированию пространства признаков, исследованию применимости различных методов машинного обучения к автоматическому обнаружению псевдонаучных публикаций и индуктивному выявлению свойств, характерных для текстов этих публикаций.

Некоторые результаты, связанные с формированием множества признаков и с построением правил для обнаружения псевдонаучных текстов, представлены в [7].

3.1. Определение псевдонауки

Псевдонаучные направления активно развиваются, появляется все больше публикаций, заявляющих о невероятных научных открытиях, которые на самом деле не имеют ничего общего с наукой.

Существует большое число различных определений псевдонауки (или лженауки), схожих в том, что основное содержание термина предполагает ошибочность позиций оппонентов рациональной науки, при этом допускается возможность их добросовестного заблуждения [64]. В настоящей работе под псевдонаукой будет пониматься любая методология или система взглядов, которая претендует на то, чтобы считаться научной, но не соблюдает принципы доказательности и аргументированности, не соответствует ни нормам научного знания, ни какой-либо области действительности, а ее предмет либо не существует, либо сфальсифицирован [65, 66].

Согласно [67], наиболее типичными признаками псевдонаучного текста являются следующие:

1) авторы чаще всего претендуют на открытие исключительной важности, решающее множество актуальных практических задач;

2) псевдонаучное открытие обычно нарушает фундаментальные законы науки;

3) авторы псевдонаучного открытия не имеют предшественников;

4) новое открытие позволяет пересмотреть все или большинство старых открытий.

Отмечаются две основные угрозы, которые несет обществу псевдонаука [67]. Во-первых, деятели псевдонауки способны завладеть на коррупционной основе бюджетными средствами под предлогом создания «прорывных технологий». Во-вторых, они часто пользуются доверием тяжело больных

людей, предлагая им «магические средства» (которые могут опираться на солидную базу промышленного производства псевдонаучной аппаратуры и на авторитет медицинских учреждений) и предотвращая от лечения, которое может предложить реальная медицина [68].

Существует небольшое количество интернет-ресурсов и периодических изданий, предоставляющих информацию о псевдонауке, с помощью которых неспециалист может ознакомиться с уже опровергнутыми учениями, чтобы не быть обманутым. Среди интернет-ресурсов можно выделить международный ресурс «RationalWiki» [66], и российский ресурс «Фрикопедия – энциклопедия лженауки» [69], которые созданы с целью систематизировать и категоризировать знания о псевдонаучных направлениях, личностях, организациях. На сайте science-freaks.livejournal.com [70] интернет-сообщество ведет обсуждение псевдонаучных работ, опубликованных в последнее время. Основным достоверным источником, детально рассматривающим вопросы, связанные с различными направлениями псевдонауки, являются периодически издаваемые бюллетени Комиссии Российской академии наук по борьбе с лженаукой и фальсификацией научных исследований под названием «В защиту науки» [71].

Псевдонаучные тексты регулярно появляются на сайтах СМИ, в журналах (обычно не рецензируемых), газетах, в патентных заявках, в материалах конференций и в других публикациях, которые не прошли должную проверку перед изданием. В связи с этим, актуальной является разработка метода, который позволит автоматически определять, является ли текст псевдонаучным, стоит ли доверять его содержанию или необходимо обратиться к ученым-специалистам за помощью.

Приведем фрагменты псевдонаучных текстов. Курсивом выделены слова и синтаксические структуры, которые позволяют определить псевдонаучный характер публикации.

Пример 1: «Эта *версия создания* возникла у меня при рассуждении о большом взрыве. *Ведь если задуматься*, откуда взялось тело, которое

взорвалось, и энергия, *что спровоцировало* взрыв одиночной, замкнутой и устойчивой системы. Это *вызвало* у меня *негодование*, и я подумал о том, что материя *могла бы возникнуть* из энергии, а энергия – в результате деятельности некоего *первоэлемента*».

Пример 2: «Неоднородность пространства и материи *подтверждается множеством научных* исследований, в том числе, и посредством *самых точных приборов*, которыми только располагает современная наука».

Тексты, которые содержат большое число подобных фрагментов, имеющих слабое отношение к науке, могут быть обнаружены путем автоматического анализа содержания текстов.

3.2. Обнаружение псевдонаучных фрагментов

3.2.1. Описание метода обнаружения псевдонаучных фрагментов

Псевдонаучные тексты характеризуются особой лексикой и особыми синтаксическими и семантическими структурами, такими как «*сенсационные материалы*», «*имеет великую историческую ценность*», «*вопрос жизни и смерти*». Можно предложить различные способы использования этих структур и лексики для выявления псевдонаучных текстов. Во-первых, применить статистические методы: выбор наиболее частых конструкций и анализ их встречаемости в тексте. Во-вторых, находить близкие тексты со схожей лексикой, как в работе [72], предлагающей метод для выявления искусственно сгенерированных научных текстов. В-третьих, используя рассмотренные синтактико-семантические структуры в качестве признаков, проводить классификацию текстов. Первый способ будет давать неудовлетворительные результаты в связи с тем, что научные и псевдонаучные тексты имеют схожую лексику, схожие словосочетания, которые, однако, употребляются в разных контекстах. Второй способ не подходит по той же причине: из-за схожести лексики, научные статьи будут похожи на псевдонаучные работы, которые близки по теме. Метод, предложенный в [72], не может быть использован в связи с тем, что псевдонаучные тексты написаны естественным языком, в них

нет того количества нехарактерных для языка синтаксических структур, которыми наполнены сгенерированные тексты. Поэтому в настоящей работе предпочтение отдается методам классификации.

Стоит отметить, что классификация по признаку «количество устойчивых общенаучных словосочетаний в тексте», представленному в разделе 2.1, не дает приемлемых результатов. Проведен эксперимент, в котором для коллекции псевдонаучных статей вычислялись значения предложенного признака. В результате установлено, что лишь около 8% псевдонаучных статей имеют заниженное (менее 20%) количество общенаучных словосочетаний и около 5% имеют низкое (менее 10%) количество (рис. 17). Остальные 87% публикаций удовлетворяют требованиям и не отличаются по этому признаку от научных статей.

Рисунок 17 – Количество общенаучных словосочетаний в псевдонаучных статьях

Если относить к научным статьям только тексты с приемлемым количеством общенаучных словосочетаний, а остальные тексты относить к псевдонаучным, то точность классификации (precision) будет приближаться к значению 1. Однако такая классификация имеет низкое значение полноты (recall) равное лишь 0,13, и, соответственно, низкую F_1-меру равную 0,23. Поэтому в работе предлагается другой способ классификации.

Перейдем к рассмотрению метода обнаружения псевдонаучных текстов, разработанного в рамках настоящей работы. Некоторые результаты, связанные с разработкой метода, представлены в [1, 8]. Поскольку псевдонаучные высказывания могут составлять лишь небольшую часть публикации, предлагается разбивать статьи на небольшие фрагменты текста, близкие по объему, и классифицировать их отдельно. Разбиение текста выполняется таким образом, чтобы фрагменты состояли из абзацев, поскольку абзац обычно несет в себе законченную мысль, и, как правило, позволяет получить представление о корректности входящих в него высказываний.

В качестве признаков классификации выбраны отдельные слова, словосочетания с синтаксическими и семантическими связями, их обобщения и триграммы, образующие речевые обороты. Обобщения строятся так, что одно из слов словосочетания заменяется названием соответствующей части речи. Так словосочетание «мистический огонь» породит обобщения «мистический <сущ.>» и «<прил.> огонь». Первое обобщение, возможно, позволит обнаружить псевдонаучные тексты, посвященные другому мистическому объекту, который не упоминался в текстах обучающей выборки. Множество признаков классификации формируется автоматически с помощью лингвистического анализатора на основе обучающей выборки, которая рассматривается ниже. В текстах обучающей выборки выявлено множество признаков классификации, среди них:

- слова: "торсионный", "гармонизировать", "чрезвычайно", "неправота";
- словосочетания с синтаксическими и семантическими связями: "повсеместное наличие", "необъяснимая аномалия", "усматривать в модели", "убедительно показать", "память воды";
- обобщения словосочетаний: "память <сущ.>", "<прил.> аномалия", "усматривать в <сущ.>";
- триграммы: "я якобы сразу", "и почти нигде", "совершенно очевидно то", "сейчас наукой доказано".

Среди признаков классификации встречаются общеупотребимые и общенаучные слова, такие как "метод", "теория", "возникновение" и др. Для придания большей значимости словам, характерным лишь для псевдонаучных текстов, вектора признаков заполняются весами слов, словосочетаний, их обобщений и триграмм, которые вычисляются для каждого фрагмента текста с помощью статистической меры *TF-IDF* [73]. Согласно такой мере больший вес получают слова с высокой частотой в пределах конкретного фрагмента текста и низкой частотой употребления в остальных текстах. Таким образом, общеупотребимые и незначимые слова, часто встречающиеся во всех текстах, будут вносить несущественный вклад при классификации.

Стоит отметить, что перед разбиением текста на фрагменты автоматически, с помощью предложенного в разделе 2.2.2 метода, выделяется и удаляется список использованных источников, чтобы оставить лишь авторский текст. Научные работы обучающей выборки представляют различные темы, некоторые из них выбираются из предметных областей близких к темам псевдонаучных статей, с целью выявления отличительных признаков у схожих по лексике текстов. Необходимость выбирать часть научных работ по заданным темам затрудняет процесс формирования обучающей выборки, но при этом классификация становится предметно независимой.

В качестве классификатора выбран метод опорных векторов (SVM – support vector machine), который хорошо зарекомендовал себя при классификации текстовой информации [74]. Эксперименты проводились с использованием алгоритма с линейной функцией ядра из открытой библиотеки для метода опорных векторов LIBSVM [75]. Общая схема метода выявления псевдонаучных публикаций представлена на рис. 18.

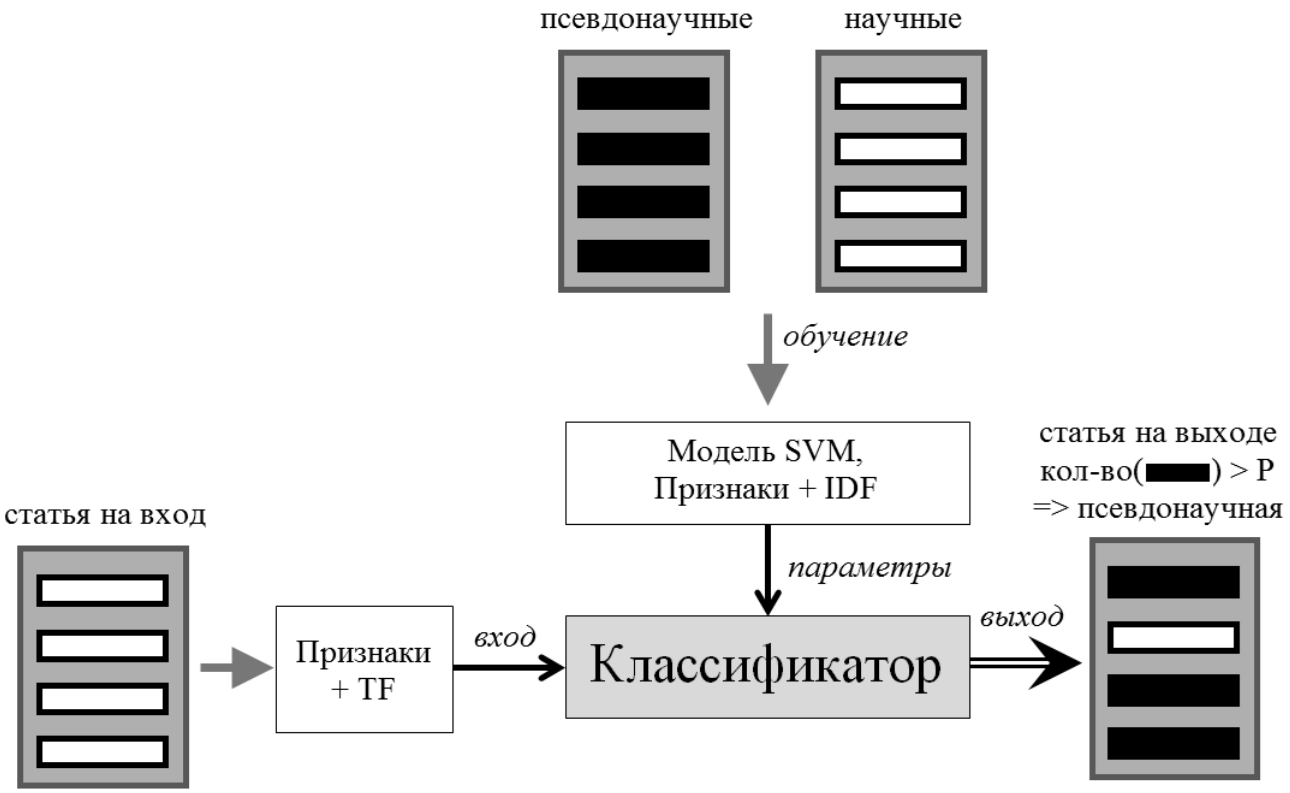


Рисунок 18 – Схема метода выявления псевдонаучных публикаций

Будем считать, что каждая публикация обучающей выборки состоит из фрагментов, принадлежащих одному и тому же классу. Так, псевдонауные тексты содержат только псевдонаучные фрагменты, научные тексты содержат только научные фрагменты, хотя такое допущение может быть причиной снижения качества обучения, если псевдонаучный текст, например, содержит фрагменты научного или нейтрального характера. Положим это же для тестовой выборки при вычислении точности и полноты классификации фрагментов. Как и в главе 2, все значения качества классификации, такие как точность, полнота и F_1-мера вычисляются согласно стандартным формулам [38].

Опишем алгоритм обнаружения псевдонаучных фрагментов в тексте. Пусть T_x – множество псевдонаучных текстов обучающей выборки, T_y – множество научных текстов обучающей выборки, таких что предметные области множества T_y включают в себя все предметные области множества T_x. Тогда обучение состоит из следующих шагов.

Алгоритм 3.1 **(алгоритм обучения классификатора для обнаружения псевдонаучных фрагментов).**

Шаг 1. С помощью метода, представленного в разделе 2.2.2, выделить в текстах списки использованных источников и удалить их, получив два новых множества $T_x^{'}$ и $T_y^{'}$, состоящих лишь из авторского текста.

Шаг 2. Разделить тексты множеств на непересекающиеся фрагменты, длина которых не превосходит, заданную величину *l* (средняя длина абзаца), так, что $T_x^{'} = \bigcup x_i$, где x_i – псевдонаучный фрагмент, аналогично $T_y^{'} = \bigcup y_i$, где y_i – научный фрагмент.

Шаг 3. Задать множество *F={∅}* – множество признаков для классификации. Выполнить синтактико-семантический анализ каждого фрагмента из множеств $T_x^{'}$ и $T_y^{'}$, расширяя множество *F* следующими признаками *t*: словами, составляющими фрагменты, словосочетаниями с синтаксическими и семантическими связями, их обобщениями и триграммами.

Шаг 4. Каждому фрагменту $d \in T_x^{'} \cup T_y^{'}$ поставить в соответствие вектор длины |*F*|, состоящий из весов признаков, вычисленных по формуле *TF-IDF*: $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$, где $tf(t,d) = \frac{n_t}{\sum_k n_k}$, где n_t – число вхождений признака *t* в фрагмент *d*, а в знаменателе – общее число признаков в данном фрагменте, $idf(t,D) = \log \frac{|D|}{|d_i \supset t|}$, где |*D*| – количество фрагментов в обучающей выборке; $|d_i \supset t|$ – количество фрагментов, в которых встречается *t* (когда $n_t \neq 0$). Задать множество $IDF = \bigcup idf(t,D)$, необходимое для последующей классификации.

Шаг 5. Выполнить обучение с помощью алгоритма SVM на полученных векторах. В результате будет получена линейная модель классификации *M*.

Наибольшей алгоритмической сложностью в предложенном методе обучения обладает алгоритм SVM: в худшем случае она равна $O(N^3)$, в среднем – $O(N^2)$, где N – число обучающих примеров [76]. Блок-схема алгоритма представлена на рис. 19.

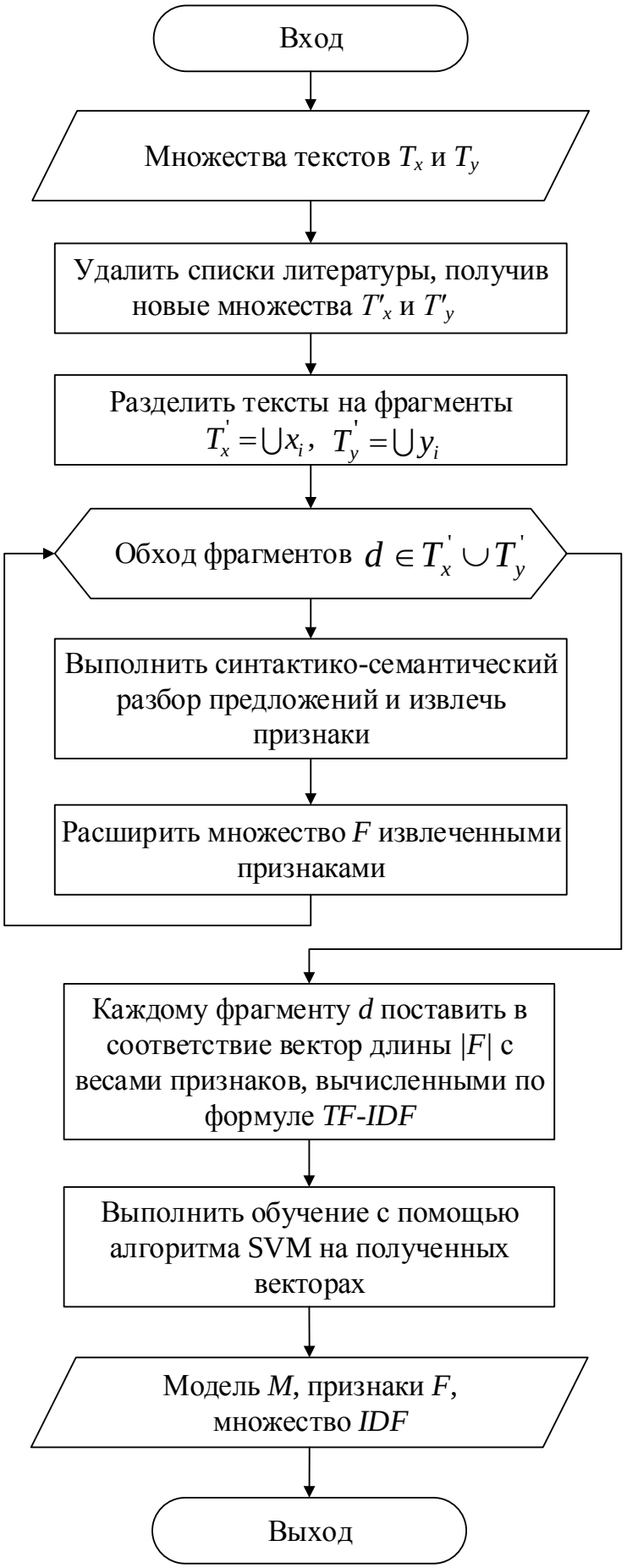


Рисунок 19 – Блок-схема алгоритма обучения SVM для классификации фрагментов

Алгоритм классификации тестового фрагмента *d* состоит в следующем.

Алгоритм 3.2 **(алгоритм классификации фрагмента текста).**

Шаг 1. Выполнить синтактико-семантический анализ предложений фрагмента *d* и извлечь признаки $F_d = \{t_i\}$.

Шаг 2. Для всех признаков, входящих в пересечение $F_d \cap F$, вычислить значение *tfidf*.

Шаг 3. Заполнить вектор длины |*F*|, используя вычисленные значения *tfidf*, и выполнить классификацию фрагмента с помощью модели *M*, полученной на этапе обучения.

Блок-схема алгоритма представлена на рис. 20.

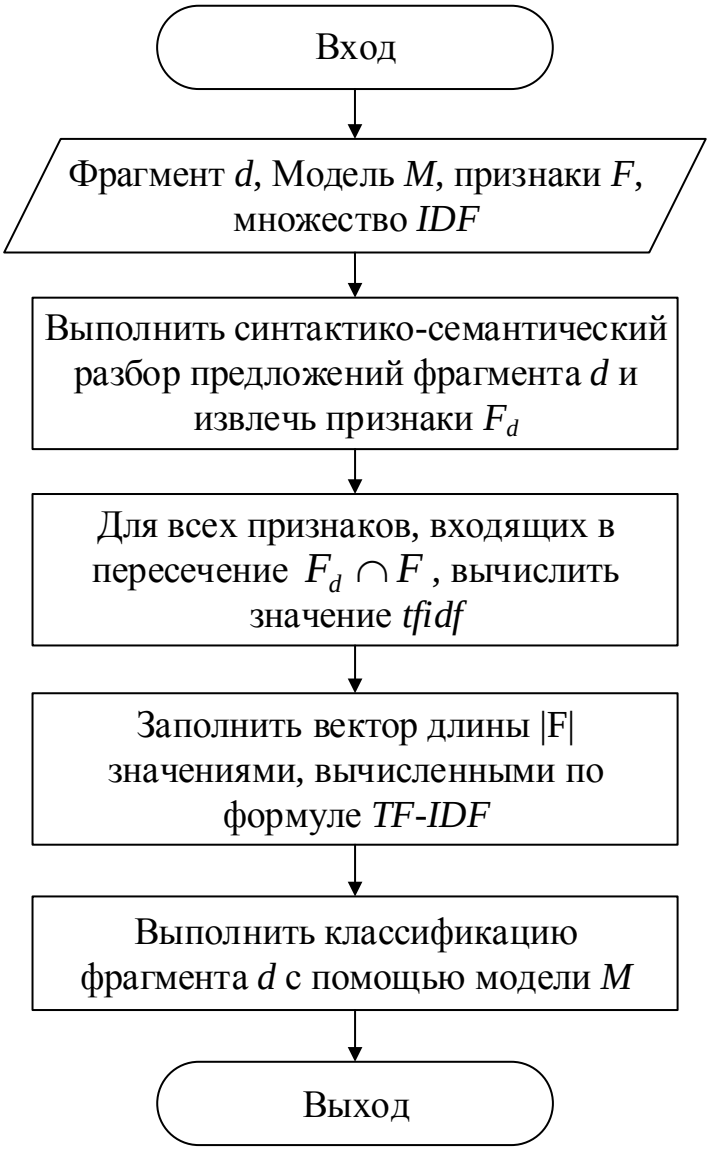


Рисунок 20 – Блок-схема алгоритма классификации фрагмента текста

3.2.2. Экспериментальная проверка метода обнаружения псевдонаучных фрагментов

Проведено два эксперимента – на небольшой выборке статей и на выборке значительно большего размера. В первом случае для обучения классификатора была сформирована выборка, в которую вошли 10 псевдонаучных статей, представленных в бюллетенях комиссии РАН по борьбе с лженаукой и 180 научных статей, отобранных экспертами из авторитетных журналов, входящих в перечень ВАК. Темы псевдонаучных статей включали следующие: необычные свойства воды, торсионные поля, телепатия. Выделено 212 фрагментов для псевдонаучных статей и 3158 фрагментов для научных, объем которых составлял около 700 символов, что соответствует среднему объему абзаца. Размерность признакового пространства составила около 350 тыс.

Сформированная тестовая выборка статей, состояла из 24 псевдонаучных и 108 научных публикаций, включающих журнальные статьи и патенты. Тестовая выборка не пересекается с обучающей выборкой. Темы псевдонаучных публикаций отличаются от тем обучающей выборки и включают: теорию эфира, последствия возможной смены полюсов Земли, парадоксы в научных теориях. Эти тексты разбиты на 374 псевдонаучных фрагмента и 2175 научных фрагментов. Для каждого фрагмента построен вектор признаков, который классифицирован с использованием модели SVM и значений IDF, полученных в качестве результата этапа обучения. В табл. 6 представлены результаты классификации фрагментов текстов. В таблице используются следующие буквенные обозначения:

- TP – истинно-псевдонаучный,
- TN – истинно-научный,
- FP – ложно-псевдонаучный,
- FN – ложно-научный.

Таблица 6 – Первый эксперимент. Результаты классификации фрагментов

Классы фрагментов публикаций		Экспертная оценка	
		псевдонаучный (374)	научный (2175)
Оценка классификатора	псевдонаучный	TP=301	FP=204
	научный	FN=73	TN=1971

Согласно табл. 6, точность классификации фрагментов (accuracy) равна

$$(TP+TN)/(TP+FN+FP+TN) = (301+1971)/(374+2175) = 0,89.$$

Тестовая выборка научных публикаций была поделена на 9 пересекающихся частей, по 24 публикации в каждой части, для вычисления точности (precision) для класса «псевдонаучный». Это сделано с целью уравнять количество работ в обоих классах и провести несколько независимых экспериментов. Значения точности для различных наборов научных статей представлены в табл. 7.

Таблица 7 – Точность классификации для различных частей тестовой выборки

№	FP	TP + FP	Точность
1	15	316	0,95
2	11	312	0,96
3	24	325	0,93
4	39	340	0,89
5	73	374	0,80
6	63	364	0,83
7	52	353	0,85
8	72	373	0,81
9	45	346	0,87
Среднее	**44**	**345**	**0,88**

Согласно табл. 7 средняя точность равна 0,88, полнота, согласно табл. 6, равна (301/374) = 0,8, средняя F_1-мера равна 0,84. В худшем случае (набор статей №5 в табл. 7) точность равна 0,8 и F_1-мера – 0,8. Несмотря на то, что значения точности и полноты не очень высокие, такого качества классификации достаточно, для того чтобы отличить научные публикации от псевдонаучных по отношению объема текста, классифицированного как псевдонаучный, к общему объему текста публикации. На рис. 21 показано,

какую часть публикации занимают фрагменты, которые классифицированы как псевдонаучные (для 24 псевдонаучных текстов и 24 научных текстов из набора №5).

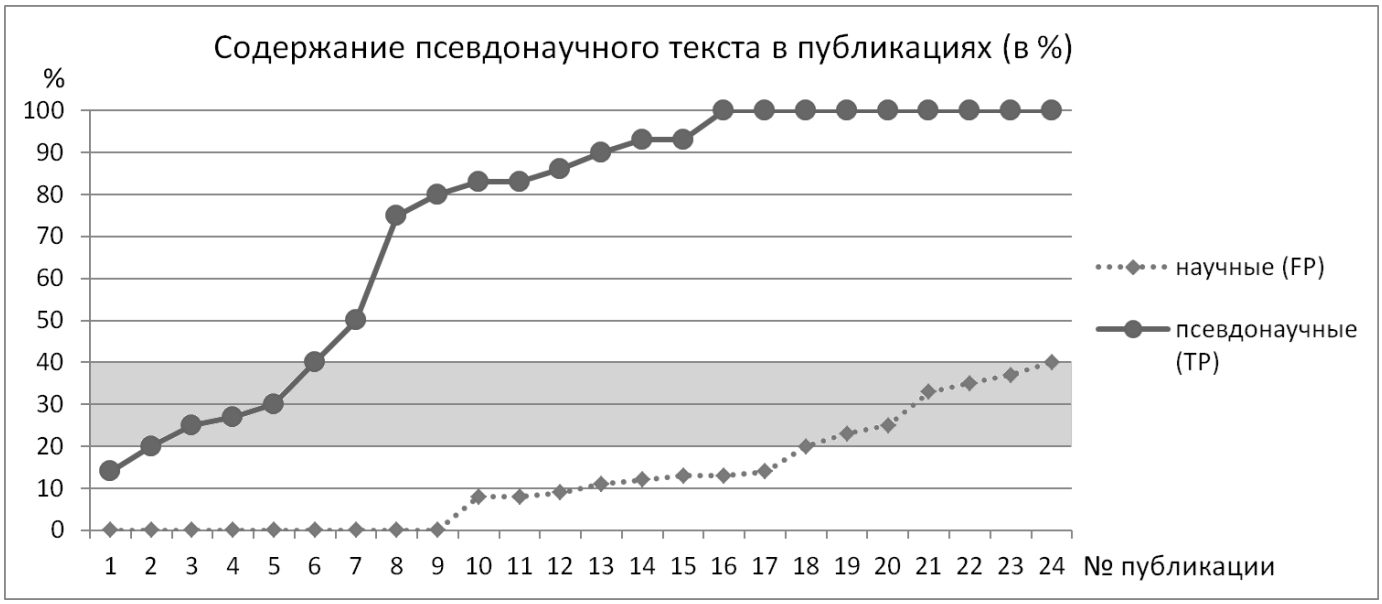


Рисунок 21 – Количество псевдонаучных фрагментов в публикациях (в %)

Примечание – статьи на графике (рис. 21) упорядочены по количеству псевдонаучных фрагментов в тексте (ось ординат).

Выделены три зоны: 0-20% – более вероятно, что публикации научные, 20-40% – как научные, так и псевдонаучные работы (требуется больше признаков для того, чтобы отличить их), 40-100% – публикации, скорее всего, псевдонаучные. Зададим порог приемлемого количества псевдонаучных фрагментов в публикации P и проведем классификацию полных текстов. Для наглядности выберем порог как среднее значение во второй зоне на рис. 21 (P = 30%) и разделим в соответствии с ним все статьи тестовой выборки на два класса. Результаты представлены в табл. 8.

Таблица 8 – Результаты классификации полных текстов

Классы публикаций		Экспертная оценка	
		псевдонаучный (24)	научный (108)
Оценка классификатора	псевдонаучный	19	7
	научный	5	101

Точность классификации полных текстов равна 19/(19+7) = 0,73, полнота равна 19/(19+5) = 0,79, F_1-мера – 0,76. В основном ошибочно отнесены к классу псевдонаучных публикаций некоторые статьи по теоретической физике, поскольку эта тема не представлена в первой обучающей выборке. Обратим внимание на то, что значение F_1-меры значительно выше полученного при классификации текстов по количеству устойчивых общенаучных словосочетаний.

Составлена коллекция псевдонаучных публикаций, содержащая более 4 тысяч статей, взятых из источников, рекомендованных комиссией РАН по борьбе с лженаукой и фальсификацией научных исследования. Второй эксперимент проводился на обучающей выборке, состоящей из 220 научных статей (3767 фрагментов) и 60 псевдонаучных статей (9001 фрагмент). Тестовая выборка состояла из 12624 авторефератов докторских диссертаций (1300K фрагментов) и 4203 статей псевдонаучных сообществ (130K фрагментов). Количество используемых признаков составило около 1 миллиона. Поскольку тестовая выборка значительно больше обучающей выборки, перекрестная проверка не проводилась.

Результаты классификации текстовых фрагментов приведены в табл. 9.

Таблица 9 – Второй эксперимент. Результаты классификации фрагментов

Классы фрагментов публикаций		Экспертная оценка	
		псевдонаучный (130K)	научный (1300K)
Оценка классификатора	псевдонаучный	83%	28%
	научный	17%	72%

Согласно табл. 9, точность классификации для класса «псевдонаучный» равна 0,75, полнота – 0,83, F_1-мера – 0,79.

Тексты тестовой выборки группировались по доле текста, которую занимают фрагменты, классифицированные как псевдонаучные. На рис. 22 и рис. 23 представлены полученные группы, каждой из которых соответствует отдельный столбик. Над столбиком указано относительное количество текстов

в этой группе, под ним – та часть текста, которую занимают псевдонаучные фрагменты.

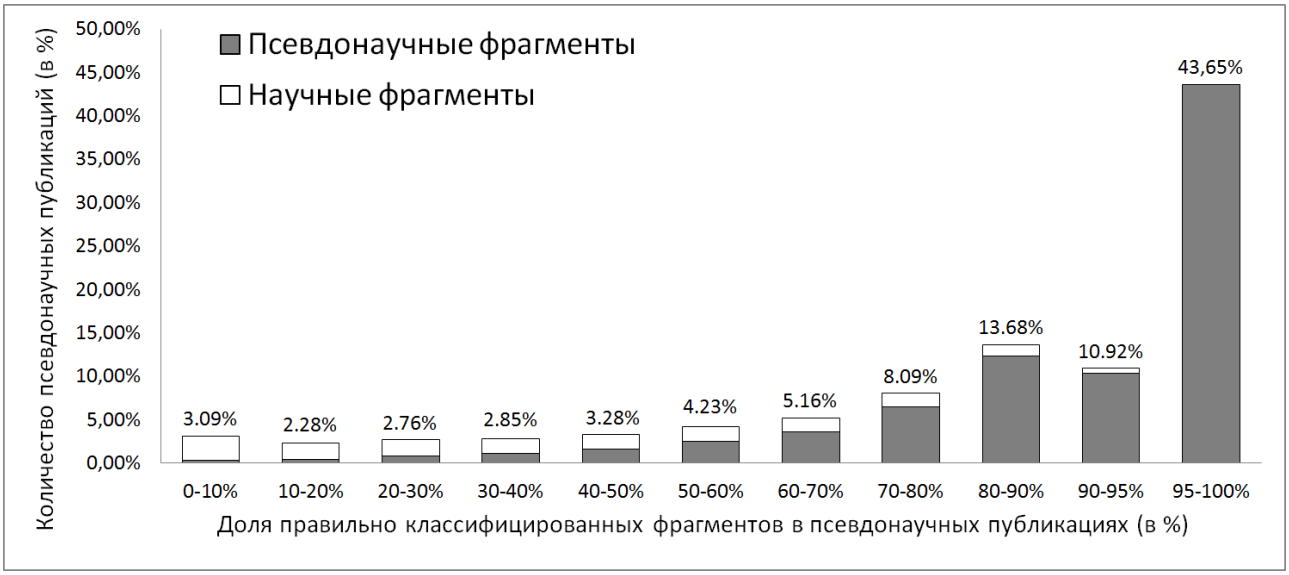


Рисунок 22 – Количество псевдонаучных фрагментов в псевдонаучных статьях (в %)

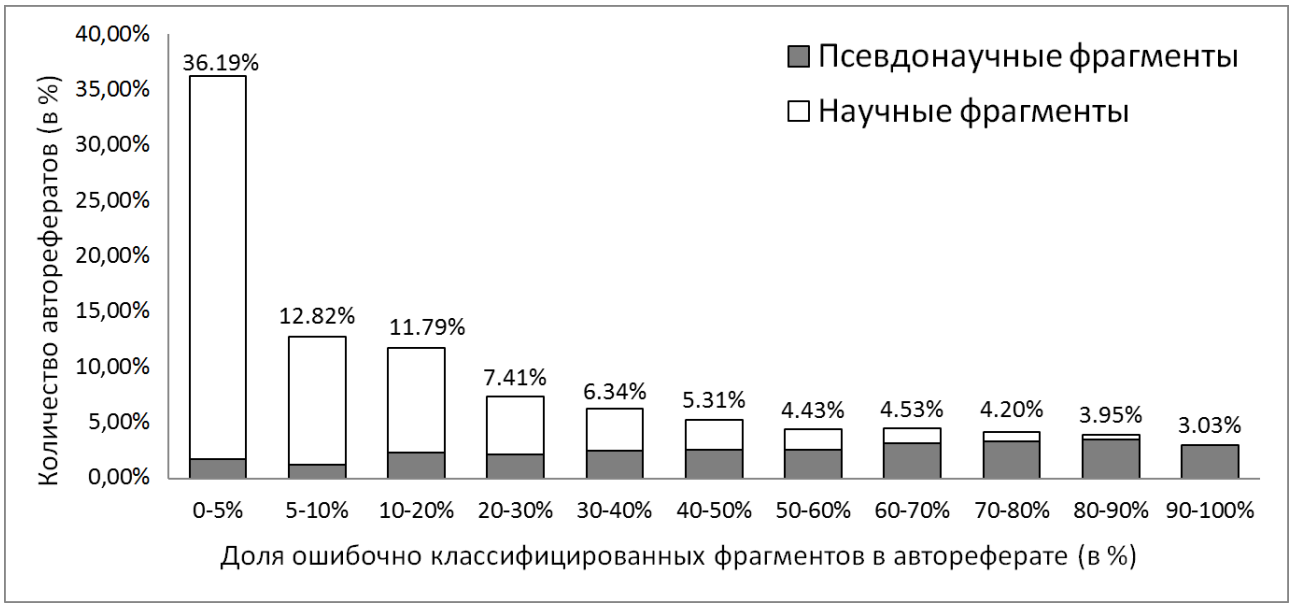


Рисунок 23 – Количество псевдонаучных фрагментов в научных публикациях (в %)

Примечание – окрашенная часть каждого столбика на рис. 22-23 соответствует диапазону, указанному под ним, который обозначает количество фрагментов в тексте, классифицированных как псевдонаучные. Так, левый столбик на рис. 23 означает, что количество авторефератов, в которых лишь от 0 до 5% фрагментов классифицировано неправильно, составляет 36,19% от всех авторефератов.

Результаты демонстрируют эффективность предложенного метода: большая часть псевдонаучных фрагментов классифицирована правильно и небольшая часть научных фрагментов классифицирована неправильно.

Согласно рис. 22, в большинстве псевдонаучных статей значительную часть текста составляют псевдонаучные фрагменты, тогда как в большинстве научных статей, согласно рис. 23, неправильно классифицированные фрагменты занимают лишь небольшой процент текста. Поэтому показатель относительного количества псевдонаучных фрагментов в тексте может быть использован как дополнительный признак, характеризующий качество текстов научной сферы.

С использованием предложенного признака (количество псевдонаучных фрагментов в тексте), на тестовой выборке проведен второй уровень классификации, где объектами являлись полные тексты. При этом порог приемлемого количества псевдонаучных фрагментов P задавался в процессе обучения так, чтобы F_1-мера на обучающей выборке была максимальна. Перекрестная проверка проведена два раза. В первом случае в качестве обучающей выборки брались 10% примеров, 90% оставалось для теста, во втором случае выполнялась классическая перекрестная проверка – 90% примеров для обучения, остальные 10% – для теста. Результаты второго уровня классификации представлены в табл. 10 и в табл. 11.

Таблица 10 – Результаты перекрестной проверки с небольшой обучающей выборкой

№	Оптимальный порог (%)	F_1-мера при обучении	F_1-мера при классификации
1	50	0,821	0,84
2	45	0,846	0,833
3	33	0,837	0,825
4	53	0,842	0,835
5	50	0,83	0,84
6	57	0,837	0,834
7	50	0,842	0,838
8	48	0,85	0,834
9	50	0,835	0,839
10	58	0,853	0,829
Среднее	**49,4**	**0,839**	**0,835**

Таблица 11 – Результаты классической перекрестной проверки

№	Оптимальный порог (%)	F_1-мера при обучении	F_1-мера при классификации
1	50	0,84	0,821
2	50	0,838	0,845
3	50	0,839	0,836
4	50	0,838	0,84
5	50	0,84	0,83
6	50	0,839	0,834
7	50	0,838	0,842
8	50	0,837	0,849
9	50	0,839	0,835
10	50	0,837	0,853
Среднее	**50**	**0,839**	**0,839**

Согласно табл. 10, средняя F_1-мера равна 0,835, максимальная F_1-мера равна 0,84, средний порог равен 49,4%. Согласно табл. 11, при классической перекрестной проверке средняя F_1-мера равна 0,839, максимальная F_1-мера равна 0,853, порог во всех случаях равен 50%. Близкие значения F_1-меры в обоих случаях, говорят о высокой предсказательной силе метода: даже при маленькой обучающей выборке классификация значительно большей тестовой выборки выполняется эффективно. На рис. 24 представлена зависимость среднего значения F_1-меры от значения параметра *P*.

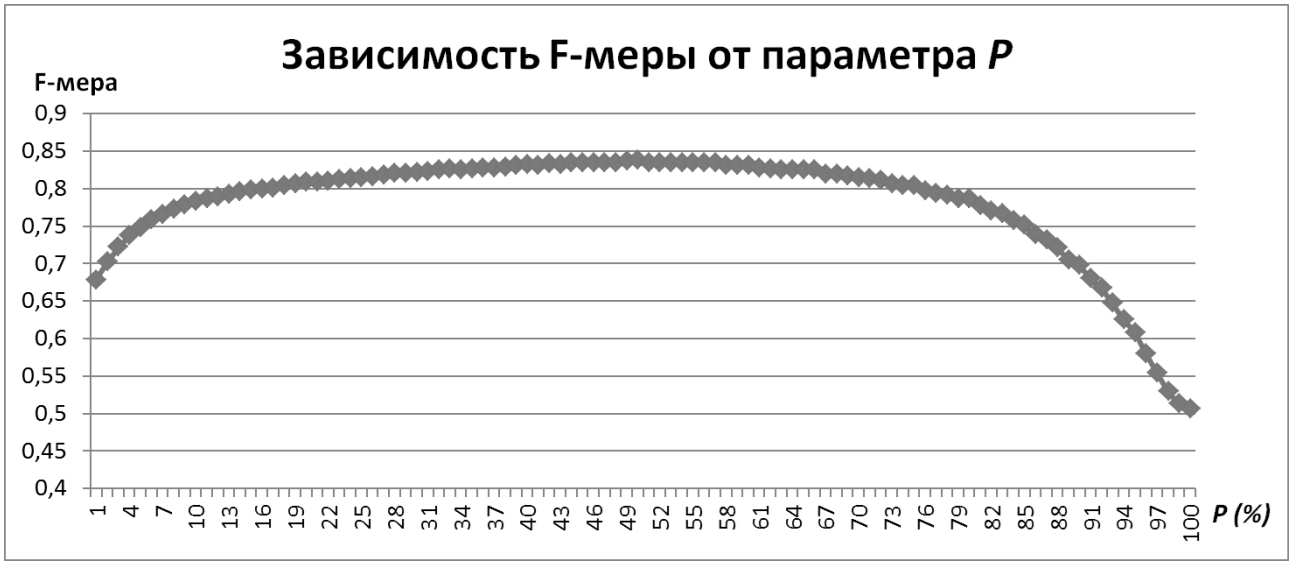


Рисунок 24 – Зависимость F_1-меры от значения параметра «приемлемое количество псевдонаучных фрагментов»

Видно, что в интервале от 16% до 75% значение F_1-меры превышает 0,8 и меняется незначительно. Это означает, что даже если всего 16% содержания статьи является сомнительным, она с высокой точностью может быть объявлена псевдонаучной. Это позволяет выявлять даже те публикации, авторы которых постарались скрыть псевдонаучную составляющую. Высокие значения F_1-меры показывают, что предложенный метод может быть использован для определения количества псевдонаучных фрагментов в тексте как еще одного признака, характеризующего качество текста. В некоторых случаях для заключения того, что текст является псевдонаучным, требуется больше признаков, в связи с чем, предлагается учитывать признаки, предложенные в

главе 2. Построим признаковое пространство с использованием всех признаков, рассмотренных в рамках настоящей работы.

3.3. Формирование признакового пространства для обнаружения псевдонаучных текстов

В главе 2 рассмотрены различные признаки, которые характеризуют качество текстов научной сферы, и предложены методы для вычисления их значений. Перечень этих признаков с множеством возможных дискретных значений приведен в табл. 12. Все эти признаки могут быть использованы при определении качества текстов путем решения задачи классификации с помощью методов машинного обучения.

Таблица 12. Перечень признаков с их значениями

	Название признака	**Значение**
1	Количество устойчивых общенаучных словосочетаний	приемлемое
		заниженное
		низкое
2	Количество необщеупотребимых словосочетаний	низкое
		завышенное
		высокое
3	Описание задачи исследования	присутствует
		вероятно, отсутствует
		отсутствует
4	Описание методов исследования	присутствует
		вероятно, отсутствует
		отсутствует
5	Описание результатов исследования	присутствует
		вероятно, отсутствует
		отсутствует
6	Выводы исследования	присутствует
		вероятно, отсутствует
		отсутствует

	Название признака	Значение
7	Список использованных источников	присутствует
		отсутствует
8	Источники в списке литературы без года опубликования	отсутствуют
		присутствуют
9	Количество источников, являющихся веб-сайтами	приемлемое
		завышенное
10	Источники в списке литературы, на которые нет ссылок в тексте	отсутствуют
		присутствуют
11	Количество источников, указанных одновременно в одной из ссылок в тексте	приемлемое
		завышенное
12	Количество цитирований работ одного автора	приемлемое
		завышенное
13	Количество псевдонаучных фрагментов	умеренное
		завышенное
		очень высокое
14	Грамматические ошибки	отсутствуют
		присутствуют
15	Нарушения согласования прилагательных с существительными в роде, числе и падеже	отсутствуют
		присутствуют
16	Нарушения подчинительной связи прилагательного	отсутствуют
		присутствуют
17	Нарушения согласования сказуемого с однородными подлежащими	отсутствуют
		присутствуют
18	Нарушения согласования причастий с определяемыми словами в причастных оборотах	отсутствуют
		присутствуют
19	Неоднозначность связи причастий с определяемыми словами в причастных оборотах	отсутствует
		присутствует
20	Употребление превосходной степени прилагательного	правильное
		неправильное

	Название признака	Значение
21	Количество нарушений синтаксической и семантической связности	низкое
		завышенное
22	Количество плеоназмов	низкое
		высокое
23	Нарушения последовательности изложения	отсутствуют
		присутствуют

Предложенное множество признаков может содержать незначимые, неинформативные с точки зрения решаемой задачи признаки, которые не влияют на результаты классификации. В связи с этим возникает задача определения таких признаков и их исключения, для решения которой применяют методы сокращения признакового пространства. Рассмотрим недавние результаты в этой области и применим один из методов для отбора признаков, наиболее значимых для определения качества текстов научной сферы.

Классические методы уменьшения признакового пространства, такие как метод сингулярного разложения и метод главных компонент, рассматриваются в [77]. Первый метод заключается в факторизации входной матрицы M, составленной из строк, которые соответствует векторам значений признаков объектов обучающей выборки. Матрица раскладывается в произведение трех: M=UDV*, где D – матрица размера m x n, у которой элементы, лежащие на главной диагонали, являются сингулярными числами (а все элементы, не лежащие на главной диагонали, являются нулевыми), а матрицы U (порядка m) и V (порядка n) – две ортогональные матрицы, состоящие из левых и правых сингулярных векторов соответственно (а V* – сопряжённо-транспонированная матрица к V). Матрица UD является новой обучающей выборкой в уменьшенном пространстве признаков. Матрица V* используется для снижения размерности тестовой выборки T так, что TV* – новая тестовая выборка с меньшим числом признаков. Второй метод – метод главных компонент –

состоит в сокращении признакового пространства путем ортогонального преобразования набора коррелированных признаков в набор главных компонент, представляющих собой линейные комбинации признаков исходного множества, для формирования которых вычисляются собственные значения ковариационной матрицы исходных данных. Оба метода работают с вещественными представлениями признаков.

В [78] для сокращения признакового пространства предлагается привлекать эксперта, который должен построить иерархическую систему. Элементами иерархии являются составные признаки, полученные путем объединения некоторых исходных признаков. Процедура агрегирования носит последовательный характер: сначала эксперт объединяет исходные признаки в группы (причем некоторые признаки могут образовывать самостоятельные группы), затем полученные группы снова объединяются в новые группы верхнего уровня и так до составления уровня, содержащего требуемое число признаков. Для каждого агрегированного признака эксперт составляет шкалы значений, которые являются некоторой комбинацией значений исходных признаков. Такой метод обычно применяется при решении задач многокритериального выбора, предполагающих обязательное участие лица, которому необходимо распределить исходную совокупность объектов (альтернатив) по категориям, исходя из своих предпочтений. Поэтому, в связи с тем, что в задаче определения качества научных текстов неизвестно, какие признаки влияют на класс объекта, и невозможно объединить признаки и построить новые шкалы значений без потери информации, рассмотренный метод не может быть применен в явном виде.

Использование аппарата нейронных сетей для отбора информативных признаков является еще одним из способов снижения размерности признакового пространства. Так в [79] рассматривается метод оценки информативности (значимости) признаков путем последовательной фиксации их значений. Идея этого метода основывается на следующем предположении: если признак является избыточным, то фиксация входной переменной,

соответствующей этому признаку, не приведет к существенному ухудшению значения целевой функции обученной нейронной сети по сравнению со значением целевой функции, полученном на исходной выборке. В качестве фиксированного значения для вещественных переменных может быть вычислено среднее значение на обучающей выборке, для качественных переменных возможна замена среднего значения на наиболее часто встречаемое значение переменной.

Существуют методы, относящиеся к классу нелинейных методов снижения размерности. Например, в [80] выполняется сокращение пространства признаков так, чтобы сохранялись попарные расстояния между объектами, то есть не менялось расположение объектов относительно друг друга. Для оценки ошибки отображения многомерных данных в пространство малой размерности может использоваться сумма квадратов разностей между расстояниями в многомерном пространстве и расстояниями в пространстве малой размерности, вычисленных попарно для всех объектов. Недостатком метода является высокая вычислительная сложность, однако существуют подходы к ее некоторому снижению [80], например, с помощью иерархической декомпозиции пространства с использованием опорных узлов, для построения которых выполняется процедура формирования вложенных кластеров объектов.

В тех задачах, когда объекты относятся к разным классам, уменьшение количества признаков часто выполняется так, чтобы классы объектов приобретали свойство высокой отделимости друг от друга. Для этого в [81] вводится специальный критерий, не зависящий от результатов классификации и представляющий собой сумму отношений сигнала к шуму. Под сигналом авторы понимают расстояние между классами, под шумом понимается внутриклассовое расстояние между объектами. Метод заключается в решении следующей задачи оптимизации: для объектов размерности p найти функционал, отображающий их в пространство размерности k ($k<p$) так, чтобы значение введенного критерия было максимальным. Решение задачи

достигается при высоких значениях сигнала и низких значениях шума, что и позволяет формировать классы объектов, далеко отстоящие друг от друга, но имеющие небольшой внутриклассовый разброс. Недостатком метода является неинтерпретируемость результатов классификации, невозможно восстановить, какие признаки были значимыми и внесли больший вклад при классификации, поскольку при уменьшении пространства теряется информация об исходных признаках.

Методы выбора признаков с помощью регуляризованных деревьев [82] и разреженных многомерных деревьев [83] представляют еще один из эффективных способов уменьшения числа используемых признаков. Главная идея заключается в наложении штрафа на выбор нового признака для разветвления в том случае, когда использование этого признака не дает высокого прироста информации по сравнению с уже достигнутым уровнем с помощью признаков, выбранных на предыдущих ветвлениях. По мнению авторов, использование моделей деревьев позволяет работать естественным образом с категориальными и числовыми переменными, с пропущенными значениями, с различным масштабированием переменных, их зависимостью и нелинейностью. Метод применим к различным техникам машинного обучения, например, таким как случайный лес (ансамбль решающих деревьев) или градиентный бустинг деревьев решений. Преимуществом метода является малая трудоемкость работы при небольших размерностях признаковых пространств.

Индуктивные методы также могут быть использованы для сокращения признакового пространства. В работе [84] предлагается выбирать признаки, наиболее часто используемые в правилах, порождаемых алгоритмом индуктивного обучения AQ (quasi-minimal algorithm) [85, 86]. Этот алгоритм заключается в построении отличительного описания класса объектов путем нахождения наборов свойств, которым удовлетворяют лишь объекты этого класса. Под свойством понимается конкретное значение или дизъюнкция значений признака. На первом шаге алгоритма случайным образом выбирается

один из объектов и строится правило, состоящее из свойств этого объекта. Затем это правило расширяется путем добавления новых значений признаков или путем удаления некоторых свойств так, чтобы правило покрывало объекты данного класса и не покрывало ни одного объекта другого класса. Если после расширения правила остались непокрытые объекты, то процедура повторяется на оставшемся множестве объектов. Поскольку выбор начального объекта и свойств для расширения правил происходит случайно, то в разных запусках алгоритма порождаются разные правила. В связи с этим в [84] предлагается проводить несколько запусков AQ с определением статистики по встречающимся в правилах свойствам и выбирать в качестве значимых признаков те, которые используются в правилах минимальных по длине и максимальных по покрытию. Показана применимость такого подхода для предварительной обработки данных для дальнейшего построения причинно-следственных связей на множестве признаков. В работах [87, 88] предлагается модификация AQ алгоритма, позволяющая избежать его многократного запуска за счет использования коэволюционного генетического алгоритма.

В настоящей работе используется описанный индуктивный метод, поскольку он позволяет выполнять снижение размерности пространства без перехода к другой системе признаков. К тому же применение этого метода делает возможным выбрать значимые признаки безотносительно конкретного метода классификации, что не ограничивает дальнейшее исследование различных методов для обнаружения псевдонаучных текстов. На экспериментальной коллекции текстов произведен многократный запуск AQ алгоритма и получено 10 наиболее информативных признаков:

- количество псевдонаучных фрагментов;
- количество нарушений синтаксической и семантической связности;
- количество плеоназмов;
- неоднозначность связи причастий с определяемыми словами в причастных оборотах;

- наличие списка использованных источников;
- источники в списке литературы, на которые нет ссылок в тексте;
- описание задачи исследования;
- описание методов исследования;
- описание выводов исследования;
- количество необщеупотребимых словосочетаний.

Выполним далее построение множества правил, содержащих признаки, характеризующие псевдонаучные тексты.

3.4. Построение множества критериев принадлежности текста множеству псевдонаучных текстов

ДСМ-метод [89, 90], представляющий собой метод индуктивного порождения гипотез, может быть использован для построения критериев в виде правил «если-то». Получение легко интерпретируемых правил является преимуществом метода, поскольку в результате классификации текста они позволяют определить, какие признаки повлияли на выбор класса.

Опишем построение правил и признаков псевдонаучного текста в терминах ДСМ-метода. Пусть $O^+ \subset O-$ положительные примеры (псевдонаучные тексты), $O^- \subset O-$ отрицательные примеры (научные тексты). Напомним, что множество $O=\{o_i\}$ – множество объектов (текстов научной сферы), где каждый объект $o = \{ p_1^{i_1},..., p_n^{i_n} \}$ представим в виде вектора значений признаков из множества $P=\{p_j\}$. Тогда гипотезы $H^+ = \bigcup p_j^{n_j}$, полученные на первом этапе ДСМ-метода и представляющие собой объединения значений некоторых признаков, и будут являться условиями принадлежности к множеству O^+: если объект o обладает признаками, которые удовлетворяют гипотезе $h \in H^+$ и не удовлетворяют ни одной гипотезе множества H^-, то $o \in O^+$. Каждая гипотеза $h \in H^+$ является интегральным признаком псевдонаучного текста.

В результате применения описанного индуктивного метода на обучающей выборке текстов получено 3 тысячи интерпретируемых правил, с которыми в дальнейшем может работать эксперт-аналитик. Ниже приведено несколько примеров:

1) Если (количество псевдонаучных фрагментов = *очень высокое*) И (описание методов исследования = *отсутствует*)
   То (публикация является псевдонаучной).
2) Если (выводы исследования = *вероятно отсутствуют*) И (количество устойчивых общенаучных словосочетаний = *заниженное*)
   То (публикация является псевдонаучной).
3) Если (количество нарушений синтаксической связности = *завышенное*)
   И (количество плеоназмов = *высокое*)
   И (список цитируемой литературы = *отсутствует*) И (выводы исследования = *вероятно отсутствуют*)
   То (публикация является псевдонаучной).
4) Если (количество псевдонаучных фрагментов = *очень высокое*) И (количество цитирований работ одного автора = *завышенное*)
   То (публикация является псевдонаучной).

Для первичной проверки применимости такого метода к обнаружению псевдонаучных текстов проведено исследование его эффективности на выборке публикаций, для которой разделение статей по классам лишь с использованием признака «количество псевдонаучных фрагментов», предложенного в разделе 3.2, не дает высокой точности. Такая выборка состояла из 2131 псевдонаучной публикации и 218 научных статей, относящихся преимущественно к техническим наукам. Выполнена процедура перекрестной поверки: выборка поделена на 10 частей с равным количеством публикаций каждого класса, на девяти из которых проводилось обучение (построение правил ДСМ-методом и нахождение оптимального порога для разделения), а на десятой части – тестирование. Разделение выборки на 10 частей выполнялось

многократно, перекрестная проверка применялась на каждом наборе с последующим усреднением результатов. На каждом этапе обучения получено более 3 тысяч правил, которые покрывали около 87% примеров из обучающей выборки псевдонаучных текстов и ни одного примера из научной выборки. Полученные правила использовались для классификации примеров тестовой выборки. Средняя точность классификации равна 0,78, средняя полнота – 0,88, F_1-мера – 0,83. На этой же выборке разделение текстов по количеству псевдонаучных фрагментов методом, представленным в разделе 3.2, дает F_1-меру в среднем равную 0,77 (данные представлены в табл. 13).

Таблица 13. Сравнение результатов классификации разными методами

	Точность	Полнота	F_1-мера
Разделение по количеству псевдонаучных фрагментов	0,68	0,89	**0,77**
ДСМ-метод (множество правил)	0,78	0,88	**0,83**

Таким образом, использование всех признаков, характеризующих качество текстов научной сферы, позволяет проводить классификацию псевдонаучных текстов с большим значением F_1-меры, что говорит о целесообразности применения разработанных в настоящей работе методов извлечения признаков.

Выполним сравнение рассмотренного индуктивного метода с методами машинного обучения.

3.5. Сравнительный анализ эффективных методов классификации

Сравним несколько методов классификации, подходящих для решения задачи обнаружения псевдонаучных текстов, а именно, рассмотренный индуктивный метод ДСМ [89], метод опорных векторов [74] и методы классификации с помощью нейронной сети (трехслойный персептрон) [91] и деревьев решений [92]. В качестве реализаций трех последних методов используются открытые библиотеки LIBSVM [75], FANN [93] и WEKA (AdaBoost) [94] соответственно.

Составим несколько различных выборок экспериментальной коллекции публикаций. В качестве первой выборки возьмем использованный в предыдущем эксперименте набор статей, который с низкой точностью разделяется по признаку «Количество псевдонаучных фрагментов». Вторая выборка состоит из тех же статей, но для обучения теперь используется десятая часть статей, остальные статьи выбраны для тестирования. Третья и четвертая выборки составлены из всех имеющихся статей, и отличаются друг от друга, как и первые две выборки, долей статей, используемых для обучения и тестирования. Данные о выборках представлены в табл. 14.

Таблица 14. Размеры экспериментальных выборок публикаций

	Количество псевдонаучных статей	Количество научных статей	Доля статей для обучения	Доля статей для теста
Выборка №1	2131	218	0,9	0,1
Выборка №2	2131	218	0,1	0,9
Выборка №3	2131	938	0,9	0,1
Выборка №4	2131	938	0,1	0,9

Статьи каждой выборки были поделены случайным образом на 10 разных частей десятью способами. Для каждого разбиения проведена перекрестная проверка с использованием всего признакового пространства. Результаты экспериментов, усредненные по всем прогонам, представлены в табл. 15-18. При вычислении значений точности и полноты учитывалось отличие количества публикаций в разных классах – использовались нормированные показатели.

Таблица 15. Результаты классификации на выборке №1 (обучающая выборка больше тестовой)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,78	0,88	**0,83**
Метод опорных векторов	0,82	0,99	**0,9**
Нейронная сеть	0,69	0,98	**0,81**
Деревья решений	0,79	0,99	**0,88**

Таблица 16. Результаты классификации на выборке №2 (обучающая выборка меньше тестовой)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,72	0,96	**0,82**
Метод опорных векторов	0,77	0,98	**0,86**
Нейронная сеть	0,65	0,99	**0,78**
Деревья решений	0,76	0,98	**0,86**

Таблица 17. Результаты классификации на выборке №3 (обучающая выборка больше тестовой)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,87	0,8	**0,83**
Метод опорных векторов	0,93	0,96	**0,95**
Нейронная сеть	0,87	0,96	**0,91**
Деревья решений	0,91	0,96	**0,93**

Таблица 18. Результаты классификации на выборке №4 (обучающая выборка меньше тестовой)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,83	0,93	**0,88**
Метод опорных векторов	0,91	0,96	**0,93**
Нейронная сеть	0,77	0,97	**0,87**
Деревья решений	0,89	0,96	**0,92**

Из таблиц видно, что во всех случаях наибольшую F_1-меру дает метод опорных векторов, и близкую к максимальной F_1-меру показывает классификация с применением деревьев решений. ДСМ-метод и нейронные сети также дают приемлемые результаты, однако для настройки нейронных сетей требуется большой набор данных, ДСМ-метод, наоборот, лучше работает при небольшом размере обучающей выборки, что говорит о его высокой предсказательной силе.

Рассмотрим результаты обнаружения псевдонаучных текстов с использованием сокращенного пространства признаков. Проведена

перекрестная проверка с использованием отобранных значимых признаков на выборках №1 и №2, представленных в табл. 14. Результаты усреднены по всем прогонам и представлены в табл. 19-20.

Таблица 19. Результаты на выборке №1 (только значимые признаки)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,92	0,75	**0,82**
Метод опорных векторов	0,82	0,99	**0,9**
Нейронная сеть	0,8	0,98	**0,88**
Деревья решений	0,79	0,99	**0,88**

Таблица 20. Результаты на выборке №2 (только значимые признаки)

	Точность	Полнота	F_1-мера
ДСМ-метод	0,81	0,92	**0,86**
Метод опорных векторов	0,77	0,99	**0,87**
Нейронная сеть	0,7	0,98	**0,82**
Деревья решений	0,77	0,98	**0,86**

Сравнение табл. 15-16 с табл. 19-20 показывает, что уменьшение признакового пространства почти не влияет на результаты классификации при использовании метода опорных векторов и деревьев решений. Точность классификации с помощью ДСМ-метода и нейронной сети увеличилась, что почти во всех случаях привело и к увеличению F_1-меры. Лишь на первой выборке в случае ДСМ-метода F_1-мера незначительно уменьшилась из-за снижения полноты. Этот и другие проведенные эксперименты показали, что лучшие результаты при применении ДСМ-метода достигаются при обучении на небольшой коллекции публикаций.

Пригодность выбранных признаков проверена на выборках равного размера. Для этого составлена контрольная выборка, содержащая 2 тыс. научных публикаций журнала «Теоретическая и математическая физика», которые процитированы, по крайней мере, один раз. Аналогично предыдущим

экспериментам выполнялся десятикратный прогон перекрестной проверки на разных разбиениях выборки. Согласно сделанным выше выводам об эффективности работы ДСМ-метода, его обучение проводилось на выборке небольшого размера. Результаты применения методов представлены в табл. 21.

Таблица 21. Результаты классификации на контрольной выборке

	Точность	Полнота	F_1-мера
ДСМ-метод	0,95	0,75	**0,84**
Метод опорных векторов	0,96	0,96	**0,96**
Нейронная сеть	0,94	0,95	**0,95**
Деревья решений	0,96	0,95	**0,95**

Из таблицы видно, что сформированное пространство признаков позволяет с высокой полнотой и точностью отделять псевдонаучные публикации от научных. ДСМ-метод показывает низкую полноту по сравнению с остальными методами, однако порождаемые им правила с высокой точностью характеризуют псевдонаучные работы, благодаря чему их можно использовать для объективного обоснования выбора класса для каждой публикации.

Результаты главы 3

В настоящей главе представлено решение задачи обнаружения псевдонаучных текстов с использованием признаков, характеризующих качество текстов научной сферы, предложенных в главе 2, и дополнительного признака, определяющего количество псевдонаучных фрагментов в тексте. Для установления значений дополнительного признака предложен метод обнаружения псевдонаучных фрагментов, заключающийся во взаимодействии лингвистических методов, используемых для выявления признаков классификации, и информационных методов, в частности, статистического метода, позволяющего устанавливать значимость признаков, и метода машинного обучения, который необходим для определения принадлежности фрагмента множеству псевдонаучных текстов. В качестве признаков классификации используются такие элементы текста, как слова, словосочетания

с синтаксическими и семантическими связями, их обобщения и триграммы. Описаны условия формирования обучающей выборки для предметно независимой классификации. Проведены эксперименты, показывающие высокую F_1-меру при классификации фрагментов авторефератов докторских диссертаций и фрагментов псевдонаучных статей.

Предложен метод распределения полных текстов по классам на основании количества псевдонаучных фрагментов в тексте. Показано, что использование методов машинного обучения и всех признаков, предложенных в настоящей работе, повышает точность и полноту обнаружения псевдонаучных текстов.

Выполнено снижение размерности признакового пространства с применением модификации индуктивного алгоритма AQ, которое показало, что наиболее информативными являются структурные признаки, признаки, связанные с лингвистическими ошибками (в том числе нарушение связности текста, некоторые нарушения согласования и лексическая избыточность), а также признак, определяющий относительное количество псевдонаучных фрагментов в тексте. С помощью ДСМ-метода сформировано множество интерпретируемых правил для обнаружения псевдонаучных текстов как на полном, так и на сокращенном пространстве признаков. Показано, что использование правил, построенных с меньшим числом признаков, повышает точность обнаружения псевдонаучных текстов.

Выполнен сравнительный анализ различных методов классификации при решении задачи распределения текстов по классам «научный/псевдонаучный». Наиболее высокие значения F_1-меры достигают метод опорных векторов и деревья решений. Нейронные сети позволяют решать задачу с высокой точностью лишь при обучении на большом числе данных, ДСМ-метод, напротив, лучше работает при небольшой обучающей выборке. При этом все методы показывают высокие значения F_1-меры, что говорит о применимости сформированного пространства признаков к автоматическому обнаружению псевдонаучных текстов.

ЗАКЛЮЧЕНИЕ

В рамках настоящей работы получены следующие основные результаты:

1. Разработан новый метод автоматического формирования общенаучного словаря устойчивых словосочетаний.

2. Разработан новый метод автоматического выявления структуры научной публикации.

3. Разработан новый метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения.

4. Впервые разработан метод автоматического выявления псевдонаучных фрагментов текстов научной сферы.

5. Сформировано множество признаков, характеризующих качество текстов научной сферы.

6. Построено множество правил для обнаружения псевдонаучных текстов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Shvets, A. A Method of Automatic Detection of Pseudoscientific Publications // Proceedings of the 7th IEEE International Conference Intelligent Systems (IS'2014 IEEE). Advances in Intelligent Systems and Computing (AISC). – Warsaw, 2015. – Vol. 2. – P. 533-539.
2. Osipov, G., Smirnov, I., Tikhomirov, I., Sochenkov, I., Shelmanov, A., and Shvets, A. Information Retrieval for R&D Support / Paltoglou, Georgios, Loizides, Fernando, Hansen, Preben (Eds.) Professional Search in the Modern World. Lecture Notes in Computer Science (LNCS). – 2014. – Vol. 8830. – P. 45-69.
3. Швец А.В., Кузнецова Ю.М., Осипов Г.С., Латышев А.В. Метод и алгоритм обнаружения признаков лингвистических дефектов в научно-технических текстах // Информационные технологии и вычислительные системы. – 2013. – № 2. – С. 79-87.
4. Кузнецова Ю.М., Осипов Г.С., Чудова Н.В., Швец А.В. Автоматическое установление соответствия статей требованиям к научным публикациям // Труды ИСА РАН. – 2012. – Т. 62. – Вып. 3. – С. 132-138.
5. Швец А.В., Смирнов И.В. Программа оценки соответствия структуры научно-технического документа предъявляемым требованиям (свидетельство № 2013613411, 2013 г.).
6. Смирнов И.В., Девяткин Д.А., Тихомиров И.А., Швец А.В. Программа выявления связей между научно-техническими документами (свидетельство № 2013613409, 2013 г.).
7. Швец А.В. Формирование признакового пространства в задачах автоматического анализа научных текстов // Труды шестой международной конференции «Системный анализ и информационные технологии» (САИТ-2015). Светлогорск, 2015. – Т. 1. – С. 222-228.

8. Швец А.В. Метод автоматического выявления псевдонаучных публикаций // Теория и практика системного анализа: Труды III Всероссийской научной конференции молодых ученых с международным участием (ТПСА'14). – Рыбинск, 2014. – Т. 2. – С. 186-193.
9. Швец А.В. Экспериментальный метод автоматического определения уровня качества научных публикаций // Труды пятой международной конференции «Системный анализ и информационные технологии» (САИТ-2013). Красноярск, 2013. – Т. 1. – С. 304-312.
10. Сенкевич М.П. Стилистика научной речи и литературное редактирование научных произведений. М.: Высшая школа, 1984. 320 с.
11. Валеева Н.Г. Жанрово-стилистическая характеристика научных текстов. Введение в переводоведение. М.: РУДН, 2006. 85 с.
12. Селезнева Н.А. Использование модальных глаголов для осуществления функций научного текста // Актуальные проблемы языкознания и литературоведения. Университетские чтения ПГЛУ. Пятигорск, 2008. [Электронный ресурс]
13. Лариохина Н. М. Вопросы синтаксиса научного стиля речи. – М.: Русский язык, 1979. – 236 с.
14. Кожина М. Н., Котюрова М. П. Изучение научного функционального стиля во второй половине XX в., "Stylistyka-VI". – Opole, 1997. – С. 145-172.
15. Bolshakova E. Common scientific lexicon for automatic discourse analysis of scientific and technical texts // International journal "Information Theories and Applications". 2008. V. 15. Pp. 189-195.
16. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication IV. A.1.a. General Principles. 2010.

17. Day R. A. The Origins of the Scientific Paper: The IMRAD Format // American Medical Writers Association Journal. 1989. V. 4. № 2. P. 16-18.
18. Свидерская И. В. Коммуникации в международном сообществе. Сайт ИФБиБТ. [Электронный ресурс] http://bio.sfu-kras.ru/?page=137 (дата обращения: 31.05.2015).
19. Sollaci L. B., Pereira M. G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey // J. Med. Libr. Assoc. 2004. V. 92. № 3. P. 364-371.
20. Розенталь Д. Э. Управление в русском языке: Словарь-справочник. Для работников печати. – М.: Книга, 1981. – 207 с.
21. Лебедева Л. Плеоназм. В кн.: Русский язык: Энциклопедия. М., 1979.
22. Ляховецкая О.Я. Виды плеонастических выражений в разноструктурных языках. В кн.: Семантические процессы и их проявление в языках разного типа. Саратов, 1985. – 129 с.
23. Бабайцева В.В, Чеснокова Л.Д. Русский язык. Теория. 5-9 классы. М.: Дрофа, 2012.
24. Steingraber, S., Jolls, C., Goldberg, D.: Guidelines for Writing Scientific Papers. Tech. rep. – 1985.
25. Szklo, M.: Quality of scientific articles. Revista de Saúde Pública 40(SPE). – pp. 30–35. – 2006.
26. Gray, C.: Quality assurance and assessment of scholarly research. Research Information Network. – p. 23. – 2010.
27. Kmet, L.M., Lee, R.C., Cook, L.S.: Standard quality assessment criteria for evaluating primary research papers from a variety of fields. No. 13, Alberta Heritage Foundation for Medical Research. – 2004.
28. Писляков. В.В. Методы оценки научного знания по показателям цитирования // Социологический журнал. – 2007. – № 1. – С. 128-140.
29. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по

искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит, 2006, с. 506-524.

30. Словарь глагольно-именных словосочетаний общенаучной речи. – М., Наука, 1973. – С. 79.
31. Словарь словосочетаний, наиболее употребительных в английской научной литературе / Сост. Э.М. Басс, Е.Ф. Дмитриева, Т.М. Эльтекова. – М.: Наука, 1968. – С. 103.
32. Nenkova, A. Automatic text understanding of content and text quality. In: Frontiers of Engineering 2011: Reports on Leading-Edge Engineering from the 2011 Symposium. – pp. 49–54. – 2012.
33. Steinberger J. Text Summarization within the LSA Framework. PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, January 2007.
34. Герасимов, С.В., Курынин, Р.В., Машечкин, И.В., Петровский, М.И., Царёв, Д.В., Шестимеров А.А. Инструментальные средства оценки качества научно-технических документов // Труды Института системного программирования РАН. – Т. 24. – С. 359-379. – 2013.
35. Rakesh P., Shivapratap G., Divya G., Soman K.P. Evaluation of SVD and NMF Methods for Latent Semantic Analysis. International Journal of Recent Trends in Engineering. – Vol. 1. – № 3. – 2009.
36. Arzucan Özgür, Dragomir R. Radev. Detecting speculations and their scopes in scientific text // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009. – Vol. 3. – pp. 1398-1407.
37. Joachims, T. Advances in Kernel Methods – Support Vector Learning, chapter Making Large-Scale SVM Learning Practical. MIT-Press. – 1999.
38. Powers, David M. W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies. – V. 2(1). – pp. 37–63. – 2011.

39. Agarwal, S., Yu, H.: Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics 25(23), 3174–3180 (2009).
40. McCallum, A. and Nigam, K. A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. The AAAI Press, Madison, Wisconsin, pp. 41–48. – 1998.
41. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28(7), 991–1000 (2012).
42. Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. Identifying sections in scientific abstracts using conditional random fields. In Proceedings of the IJCNLP. – 2008.
43. Waard, A., Buitelaar, P., Eigner, T. Identifying the epistemic value of discourse segments in biology texts. Proceedings of the Eighth International Conference on Computational Semantics, IWCS-8 '09. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. pp. 351–354.
44. Liakata, M., Thompson, P., de Waard, A., Nawaz, R., Maat, H.P., Ananiadou, S. A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction // Proceedings of the Workshop on Detecting Structure in Scholorly Discourse (DSSD). – 2012. – pp. 37–46.
45. Сарыбеков, М., Сыдыкназаров, М. Словарь науки. Общенаучные термины и определения, науковедческие понятия и категории: Учебное пособие. Издание 2-ое, доп. и перераб. - Алматы: ТРИУМФ-Т, 2008. - 504 с.
46. Рябцева Н.К. Научная речь на английском языке. Руководство по научному изложению. Словарь оборотов и сочетаемости общенаучной лексики / Н. К. Рябцева. – 1999.
47. Гельбух, А. Ф., Сидоров, Г. О., Эрнандес-Рубио Э. Словари сочетаемости слов: какой метод составления лучше? // Труды международной конференции "Диалог 2004". – 2004.

48. Захаров В.Н., Хорошилов А.А., Хорошилов А.А. Опыт создания кластеров документов на основе метода определения их тематического подобия // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL-2014, Дубна, 2014. – С. 322-328.
49. Никитин Ю.В., Хорошилов А.А., Хорошилов А.А. Методы автоматического построения формализованного представления содержания материалов электронных средств массовых коммуникаций для решения задачи мониторинга и оценки деятельности органов власти // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL-2014, Дубна, 2014. – С. 145-152.
50. Формирование модели сочетаемости слов русского языка и исследование ее свойств / Э.С. Клышинский и др. // Препринты ИПМ им. М.В. Келдыша. 2013. № 41. 23 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-41 (дата обращения: 31.05.2015).
51. Арефьев Н.В. Методы построения и использования компьютерных словарей сочетаемости для синтаксических анализаторов русскоязычных текстов [Текст] : авторефер. дис. … канд. физико-мат. наук : 05.13.11 / Н.В. Арефьев. – М., 2012. – 22 с.
52. Сокирко А. В. Семантические словари в автоматической обработке текста: По материалам системы ДИАЛИНГ: Дисс. … канд. тех. наук. Москва, 2001. 120 с.
53. Осипов Г. С., Смирнов И. В., Тихомиров И. А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. – 2008. – № 2. – С. 3-10.

54. Сулейманов Д.Ш., Гатиатуллин А.Р. Модель многословных конструкций татарского языка: аналитические формы // Казанская наука. – 2012. – № 12. – С. 220-223.
55. Куршев Е. П., Сулейманова Е. А., Трофимов И. В. Роль знаний в системах извлечения информации из текстов // Программные системы: теория и приложения. – 2012. – № 3(12). – С. 57–70.
56. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997. – 112 с.
57. Национальный корпус русского языка [Электронный ресурс]: URL: http://ruscorpora.ru/ (дата обращения: 31.05.2015).
58. Manning C. D., Raghavan P., Schütze M. Introduction to Information Retrieval // Cambridge University Press, 2008. P. 240.
59. Councill, I.G., Giles, C.L., Kan, M.Y. Parscit: an open-source crf reference string parsing package. In: Proceedings of LREC. – Vol. 28. – pp. 661-667. – 2008.
60. Kern, R., Kampfl, S. Extraction of references using layout and formatting information from scientific articles. D-Lib Magazine. – Vol.19. – № 9. – 2013.
61. Академия наук СССР институт русского языка «Русская грамматика». – М.: Наука, 1980. – Т. 2. – 720 с.
62. Розенталь Д.Э., Джанджакова Е.В., Кабанова Н.П. Справочник по правописанию, произношению, литературному редактированию. М.: ЧеРо, 1999.
63. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Белгород: БГТУ, 2012. – Т. 4. – С. 100-108.

64. Александров Е. Б. Проблемы экспансии лженауки / Бюллетень «В защиту науки». – № 1. – С.14-29. – 2006.
65. Кувакин В.А. Интернет пресс-конференция члена Комиссии РАН по борьбе с лженаукой и фальсификацией научных исследований // Lenta.ru, 04.05.2010 г.
66. RationalWiki [Электронный ресурс]: URL: http://rationalwiki.org/ (дата обращения: 31.05.2015).
67. Александров Е. Б. Ответы на вопросы граждан о лженауке / Бюллетень «В защиту науки». – 2011. – № 8.
68. Гительзон И.И. Нужна государственная защита народа от натиска лжемедицины / Бюллетень «В защиту науки». – № 2. – С. 52-55. – 2007.
69. Фрикопедия – энциклопедия лженауки [Электронный ресурс]: URL: http://freakopedia.ru/ (дата обращения: 31.05.2015).
70. Science-freaks [Электронный ресурс]: URL: http://science-freaks.livejournal.com/ (дата обращения: 31.05.2015).
71. Бюллетень «В защиту науки». – 2013. №12. – С. 83.
72. Labbé, C., Labbé, D. Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? Scientometrics. 94(1), pp. 379-396 (2013).
73. Salton, G., Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 24(5), 513-523 (1988).
74. Cortes, C., Vapnik, V. Support-vector networks. Machine Learning. 20(3), 273 (1995).
75. LIBSVM – A Library for Support Vector Machines, http://w.csie.org/~cjlin/libsvm.
76. Léon Bottou and Chih-Jen Lin: Support Vector Machine Solvers, in Large Scale Kernel Machines, Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston editors, 1–28, MIT Press, Cambridge, MA., 2007.

77. Zaoralek, L., Peterek, T., Dohnalek, P., Gajdos, P. Comparison of Feature Reduction Methods in the Task of Arrhythmia Classification // Proceedings of the 5th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA). – 2014. – V. 303. – P. 375-382.
78. Петровский, А.Б., Лобанов, В.Н. Многокритериальный выбор в пространстве признаков большой размерности: мультиметодная технология ПАКС-М // Искусственный интеллект и принятие решений. – 2014. – № 3. – С. 92-104.
79. Волкова, С.С. Отбор информативных признаков с помощью нейронных сетей // Актуальные проблемы авиации и космонавтики. – 2014. – Т. 1. – № 10. – С. 287-288.
80. Мясников, Е.В. Выбор способа декомпозиции пространства признаков для нелинейного снижения размерности // Компьютерная оптика. – 2014. – Т. 38. – №4. – С. 790-797.
81. Yu, Y., McKelvey, T., Kung, S. Y. Feature Reduction Based on Sum-of-SNR (SoSNR) Optimization // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2014. – P. 6806-6810.
82. Deng, H., Runger, G. Feature selection via regularized trees // Proceedings of the International Joint Conference on Neural Networks (IJCNN). – 2012. – P. 1-8.
83. Deng, H., Baydogan, M.G., Runger, G. SMT: Sparse multivariate tree // Statistical Analysis and Data Mining. – 2014. – V. 7. – P. 53-69.
84. Panov A. I. Extraction of cause-effect relationships from psychological test data using logical methods // Scientific and Technical Information Processing. – 2014. – Vol. 41. – № 5. – P. 1-8.
85. Michalski R.S. AQVAL/1-Computer Implementation of Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition // Proc. Of the First Int. Joint Conf. on Pattern Recognition. Washington, DS, 1973. P. 3-17.

86. The aq21 natural induction program for pattern discovery: Initial version and its novel features / Janusz Wojtusiak, Ryszard S. Michalski, Kenneth A. Kaufman, Jaroslaw Pietrzykowski // ICTAI. – 2006. – P. 523-526.
87. Панов А.И., Швец А.В. Эволюционный метод покрытий для составления базы фактов ДСМ-метода // Четырнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ–2014 (24–27 сентября 2014г., г. Казань, Россия): Труды конференции. – Т. 2. – Казань : Издательство КФУ, 2014. – С. 323–330.
88. Lupatov A. Yu et al. Assessment of Dendritic Cell Therapy Effectiveness Based on the Feature Extraction from Scientific Publications / Lupatov A. Yu., Panov A. I., Suvorov R. E., Shvets A. V., Yarygin K. N., Volkova G. D. // Труды конференции International Conference on Pattern Recognition Applications and Methods. - Scitepress. - 2015. – Т. 2. – pp. 270-276.
89. Финн. В.К. ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных системах. // НТИ, № 11, 2000.
90. Автоматическое порождение гипотез в интеллектуальных системах / сост. Е. С. Панкратова, В. К. Финн. – М.: ЛИБРОКОМ, 2009. – 528 с.
91. Hertz, J., Palmer, R. G., Krogh. A. S. Introduction to the theory of neural computation, Perseus Books. – 1990. – 327 p.
92. Murthy S. Automatic construction of decision trees from data: A multidisciplinary survey. Data Mining and Knowledge Discovery. – 1998. – V. 2(4). – pp. 345-389.
93. FANN – Fast Artificial Neural Network Library, http://leenissen.dk/fann/wp/.
94. WEKA 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.

ПРИЛОЖЕНИЕ 1

Описание программных модулей анализа качества текста в системе «Exactus Expert»

Аннотация

Приложение содержит описание двух основных программных модулей анализа качества текста, внедренных в систему интеллектуального поиска и анализа научных публикаций «Exactus Expert» (http://expert.exactus.ru). Описаны функциональные назначения программных модулей, их логика, включающая описание структуры программного обеспечения и его составных частей (программ), описание функций составных частей, связей между ними, сведения о языке программирования, описание входных и выходных данных составных частей.

В заключении, на рис.1-3, приведены снимки системы и примеры отчетов, получаемых в результате работы модулей.

1. Программный модуль анализа словосочетаний и структуры текста

1.1 Общие сведения

Программный модуль анализа словосочетаний и структуры текста входит в состав системы интеллектуального поиска и анализа научных публикаций «Exactus Expert». Программный модуль анализа словосочетаний и структуры текста предназначен для использования в операционных системах семейства Linux (версии ядра 3.1 или более поздних), в частности, в дистрибутивах Debian GNU/Linux 6 и Debian GNU/Linux 7.

При разработке программного модуля анализа словосочетаний и структуры текста использованы следующие языки программирования высокого уровня:

C++ спецификации стандарт ISO/IEC 14882:2011, GNU Coding Standards.

1.2 Функциональное назначение

Программный модуль анализа словосочетаний и структуры текста определяет количество устойчивых общенаучных словосочетаний в тексте, наличие в тексте структурных разделов, таких как постановка проблемы, методы, результаты, выводы, определяет наличие в тексте списка использованных источников и вычисляет степень соответствия структуры формальным требованиям.

1.3 Структура программы с описанием функций составных частей и связи между ними

Программный модуль анализа структуры текста содержит подпрограмму анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp.

Подпрограмма анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp

Подпрограмма анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp реализует следующие функции:

- определяет в тексте количество устойчивых общенаучных словосочетаний;
- определяет наличие в тексте структурных разделов, таких как постановка проблемы, методы, выводы;
- определяет наличие в тексте списка использованных источников;
- вычисляет степень соответствия структуры формальным требованиям.

Для связи данной подпрограммы с пользователем используется протокол CGI. Выводимый результат имеет формат XML.

Подпрограмма qualityprocessor.cpp включает в себя следующие элементы:

1) набор классов для хранения загруженных устойчивых словосочетаний общенаучного словаря, синтаксических и семантических маркеров и результатов анализа текстов научной сферы;

2) набор функций для анализа текста и проверки соответствия структуры формальным требованиям.

Набор классов для хранения загруженных устойчивых словосочетаний общенаучного словаря, синтаксических и семантических маркеров и результатов анализа текстов научной сферы

Класс InitData предназначен для хранения элементов загруженных устойчивых словосочетаний общенаучного словаря и синтаксических и семантических маркеров, необходимых для анализа структуры текста.

```
class InitData{
public:
    google::sparse_hash_map<string, double>  sci_speech;

    google::sparse_hash_map<string, double> phr_intro;
    google::sparse_hash_map<string, double> phr_met;
    google::sparse_hash_map<string, double> phr_res;
    google::sparse_hash_map<string, double> phr_conc;
};
```

Класс QualityResults хранит результаты анализа текстов научной сферы.

```
class QualityResults{
public:
    double m_generalLexics;
    double m_introprob;
    double m_metprob;
    double m_resprob;
    double m_concprob;
    int m_litr;
    int m_scienceEstimate;
};
```

Набор функций для анализа текста и проверки соответствия структуры формальным требованиям

Функция process производит вызов функций анализа текста и на основании полученных результатов вычисляет степень соответствия структуры формальным требованиям.

```
ResultCode process(const std::string& text,
     lng::syntax::CText& synText,
     lng::semantics::CText& semText,
     lng::semantics::CRoleText& roleText,
     QualityResults& qualityResults);
```

Функция BuildCombinations выполняет извлечение из текста словосочетаний с синтаксическими и семантическими связями.

```
int BuildCombinations(lng::syntax::CText& syn_text,
     lng::semantics::CText& sem_text,
     lng::semantics::CRoleText& role_text,
     google::sparse_hash_map<string, int> &combs);
```

Функция CountStructureCombs выполняет формирование оценки наличия структурного раздела.

```
double CountStructureCombs(google::sparse_hash_map<string, int> &text_combs,
     google::sparse_hash_map<string, double> &init_combs,
     double eps);
```

Функция CountScienceCombs выполняет определение относительного количества устойчивых общенаучных словосочетаний в тексте.

```
double CountScienceCombs(google::sparse_hash_map<string, int> &text_combs,
     google::sparse_hash_map<string, double> &sci_speech);
```

Функция ReferencesExtraction извлекает из текста и структурирует список использованных источников

int ReferencesExtraction(const std::string& text);

1.4 Связи программы с другими программами

Программный модуль анализа словосочетаний и структуры текста использует модуль лингвистического анализа текстов, который передает в качестве результатов структуры, содержащие лексические, морфологические, синтаксические и семантические характеристики текста, необходимые для извлечения словосочетаний и выявления структуры текста.

Функциональность модуля анализа словосочетаний и структуры текста доступна непосредственно через программные вызовы. Модуль взаимодействует с другими частями системы «Exactus Expert» с помощью протокола CGI.

Для работы подпрограмм необходимы следующие свободно распространяемые библиотеки и утилиты, связываемые с данной подпрограммой на уровне исходного кода:

а) стандартная библиотека языка C++;

б) набор библиотек boost, содержащий высокоуровневые обёртки для утилитных функций общего назначения.

1.5 Вызов и загрузка

Программный модуль анализа словосочетаний и структуры текста является исполняемым модулем, взаимодействующим с вызывающей стороной по протоколу CGI. В общем виде процесс вызова подпрограммы выполняется с помощью элементов веб-интерфейса пользователя.

В результате работы подпрограмма выведет на экран данные в формате HTML.

1.6 Входные и выходные данные

Входными данными программного модуля анализа словосочетаний и структуры текста являются текст научной сферы, устойчивые словосочетания общенаучного словаря, синтаксические и семантические маркеры и результаты лингвистического анализа.

Выходными данными программного модуля анализа словосочетаний и структуры текста являются степень соответствия структуры формальным требованиям, информация о наличии или отсутствии каждого из структурных разделов и количество устойчивых словосочетаний общенаучного словаря в тексте.

Подпрограмма анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp

Табл. 1 – Входные данные подпрограммы анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp

Название	Характеристика
Text	Текст научной сферы.
LinguisticsResults	Результаты лингвистического анализа.
Science_speech	Устойчивые общенаучные словосочетания
IntroMarkers	Маркеры раздела «Постановка проблемы»
MetMarkers	Маркеры раздела «Методы»
ResMarkers	Маркеры раздела «Результаты»
ConcMarkers	Маркеры раздела «Выводы»

Табл. 2 – Выходные данные подпрограммы анализа словосочетаний и структуры текстов научной сферы на основе синтактико-семантического анализа и шаблонных выражений qualityprocessor.cpp

Название	Характеристика
m_scienceEstimate;	Натуральное число из интервала от нуля до пяти – степень соответствия структуры формальным требованиям.
m_introprob;	Дробное число из интервала от нуля до единицы – оценка наличия раздела «Постановка проблемы».
m_metprob	Дробное число из интервала от нуля до единицы – оценка наличия раздела «Методы».
m_resprob	Дробное число из интервала от нуля до единицы – оценка наличия раздела «Результаты».
m_concprob	Дробное число из интервала от нуля до единицы – вероятность наличия раздела «Выводы».
m_litr	Определяет характеристики списка использованных источников.
m_generalLexics;	Определяет относительное количество устойчивых общенаучных словосочетаний

2. Программный модуль обнаружения лингвистических ошибок

2.1 Общие сведения

Программный модуль обнаружения лингвистических ошибок входит в состав системы интеллектуального поиска и анализа научных публикаций «Exactus Expert». Программный модуль обнаружения лингвистических ошибок предназначен для использования в операционных системах семейства Linux (версии ядра 3.1 или более поздних), в частности, в дистрибутивах Debian GNU/Linux 6 и Debian GNU/Linux 7.

При разработке программного модуля обнаружения лингвистических ошибок использованы следующие языки программирования высокого уровня:

C++ спецификации стандарт ISO/IEC 14882:2011, GNU Coding Standards.

2.2 Функциональное назначение

Программный модуль обнаружения лингвистических ошибок выполняет выявление нарушений падежного согласования, нарушений согласования однородных подлежащих и управляющего слова, нарушений синтаксической и семантической связности в тексте, определение количества плеоназмов и вычисление интегральной оценки количества нарушений.

2.3 Структура программы с описанием функций составных частей и связи между ними

Программный модуль обнаружения лингвистических ошибок содержит подпрограмму выявления нарушений в текстах научной сферы defects.hpp.

Подпрограмма выявления нарушений в текстах научной сферы defects.hpp

Подпрограмма выявления нарушений в текстах научной сферы defects.hpp реализует следующие функции:

- определяет количество нарушений падежного согласования;
- выявляет нарушения согласования однородных существительных и управляющего слова;
- вычисляет степень нарушения синтаксической и семантической связности в тексте;
- определяет количество плеоназмов;
- вычисляет интегральную оценку количества нарушений.

Для связи данной подпрограммы с пользователем используется протокол CGI. Выводимый результат имеет формат XML.

Подпрограмма defects.hpp включает в себя следующие элементы:

1) структуры для хранения данных о выявленных нарушениях, интегральной оценки количества нарушений и предложений с отмеченными нарушениями;
2) набор функций для анализа текстов и выявления лингвистических ошибок.

Структуры для хранения данных о выявленных нарушениях, интегральной оценки количества нарушений и предложений с отмеченными нарушениями.

Структура данных estimation_t предназначена для хранения результатов выявления нарушений в тексте.

```
struct estimation_t{
   bool m_ruslangPaper; // язык текста
   double m_defectRosental; // Количество нарушений падежного согласования
   bool m_defectConnection; // Количество нарушений синтаксической связности
   double m_defectPredicate; // Количество нарушений согласования однородных
   существительных и управляющего слова
   bool m_defectPleonazm; // Количество плеоназмов

   double m_commonDefectValue;//Интегральная оценка количества нарушений
   vector<string> m_defect_captions;
   vector<pair<int, string> > m_colored_defects;
};
```

Структура defectSentence предназначена для хранения предложений с отмеченными в них нарушениями.

```
struct defectSentence{
   int sent_begin, sent_end;
   vector<pair<int, int> > defect_words;
   defectSentence(int _sent_begin, int _sent_end, vector<pair<int, int> > _defect_words){
        sent_begin=_sent_begin;
        sent_end=_sent_end;
        defect_words=_defect_words;
    };
};
```

Набор функций для анализа текстов и выявления лингвистических ошибок

Функция process производит вызов функций выявления нарушений и формирование результатов для выдачи пользователю.

```
void process(exlp::lingua_data_t& lingData,
             defective::estimation_t& estim,
             bool detailed);
```

Функция CheckRussian предназначена для проверки того, что текст написан на русском языке.

```
int CheckRussian(const std::string& text);
```

Функция DefectRosental выявляет нарушения падежного согласования.

```
int DefectRosental(syntax::CText& syn_text,
             bool detailed,
             int defect_type);
```

Функция DefectConnection позволяет вычислить степень нарушения синтаксической и семантической связности в тексте.

```
int DefectConnection(syntax::CText& syn_text,
                     semantics::CText& sem_text,
                     semantics::CRoleText& role_text,
                     bool detailed,
                     int defect_type);
```

Функция DefectPredicate выявляет нарушения согласования однородных существительных и управляющего слова.

```
int DefectPredicate(syntax::CText& syn_text,
                    bool detailed,
                    int defect_type);
```

Функция DefectPleonazm определяет уровень содержания плеоназмов.

```
int DefectPleonazm(syntax::CText& syn_text,
                   bool detailed,
                   int defect_type);
```

Функция getColoredSentence предназначена для расстановки внутри предложений тегов, позволяющих отметить нарушения.

```
string getColoredSentence(const
                   syntax::ComplexSentence::WordList& listwords,
                   vector<pair<int, int> > &bold);
```

2.4 Связи программы с другими программами

Программный модуль обнаружения лингвистических ошибок использует модуль лингвистического анализа текстов, который передает в качестве результатов структуры, содержащие лексические, морфологические, синтаксические и семантические характеристики текста, необходимые для выявления нарушений.

Функциональность модуля обнаружения лингвистических ошибок доступна непосредственно через программные вызовы. Модуль взаимодействует с другими частями системы «Exactus Expert» с помощью протокола CGI.

Для работы подпрограмм необходимы следующие свободно распространяемые библиотеки и утилиты, связываемые с данной подпрограммой на уровне исходного кода:

а) стандартная библиотека языка C++;

б) набор библиотек boost, содержащий высокоуровневые обёртки для утилитных функций общего назначения.

2.5 Вызов и загрузка

Программный модуль обнаружения лингвистических ошибок является исполняемым модулем, взаимодействующим с вызывающей стороной по протоколу CGI. В общем виде процесс вызова подпрограммы выполняется с помощью элементов веб-интерфейса пользователя.

В результате работы подпрограмма выведет на экран данные в формате HTML.

2.6 Входные и выходные данные

Входными данными программного модуля обнаружения лингвистических ошибок являются текст научной сферы и результаты лингвистического анализа.

Выходными данными программного модуля обнаружения лингвистических ошибок являются интегральная оценка количества нарушений в тексте, степень нарушений отдельных правил и предложения с выделенными нарушениями.

Подпрограмма выявления нарушений в текстах научной сферы defects.hpp

Табл. 3 – Входные данные подпрограммы выявления нарушений в текстах научной сферы defects.hpp

Название	Характеристика
Text	Текст научной сферы.
LinguisticsResults	Результаты лингвистического анализа.

Табл. 4 – Выходные данные подпрограммы выявления нарушений в текстах научной сферы defects.hpp

Название	Характеристика
m_ruslangPaper	Определяет является ли язык текста русским.
m_defectRosental	Натуральное число – количество нарушений падежного согласования.
m_defectConnection	Степень нарушения синтаксической и семантической связности.
m_defectPredicate	Натуральное число – количество нарушений согласования однородных существительных и управляющего слова.
m_defectPleonazm	Определяет количество плеоназмов.
m_commonDefectValue	Дробное число – общая оценка количества нарушений в тексте.
m_defect_captions	Список строк – названия выявленных нарушений.
m_colored_defects	Список строк – предложения с отмеченными нарушениями.

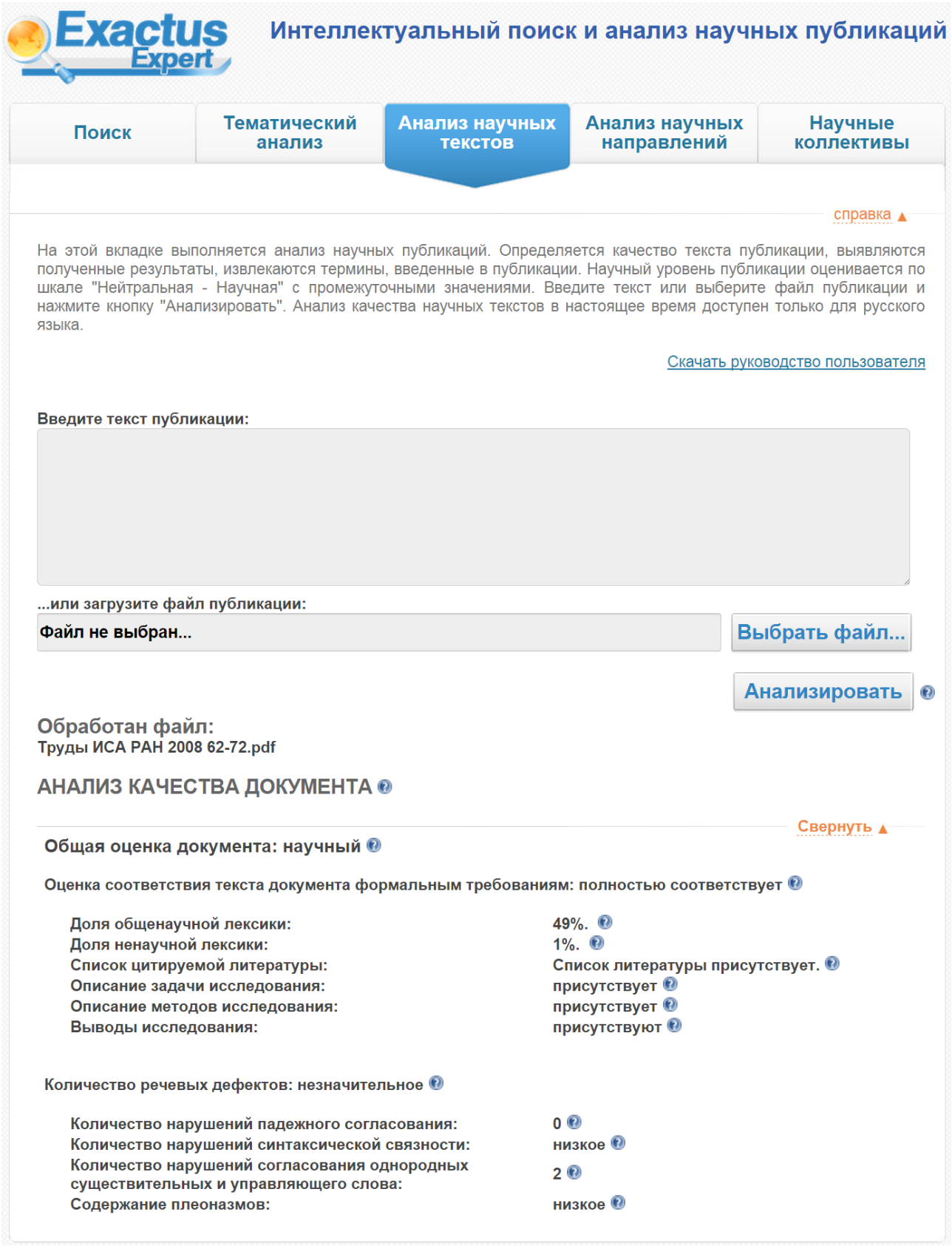


Рис.1 – Снимок системы Exactus Expert (http://expert.exactus.ru).
Пример отчета для публикации с высоким качеством текста

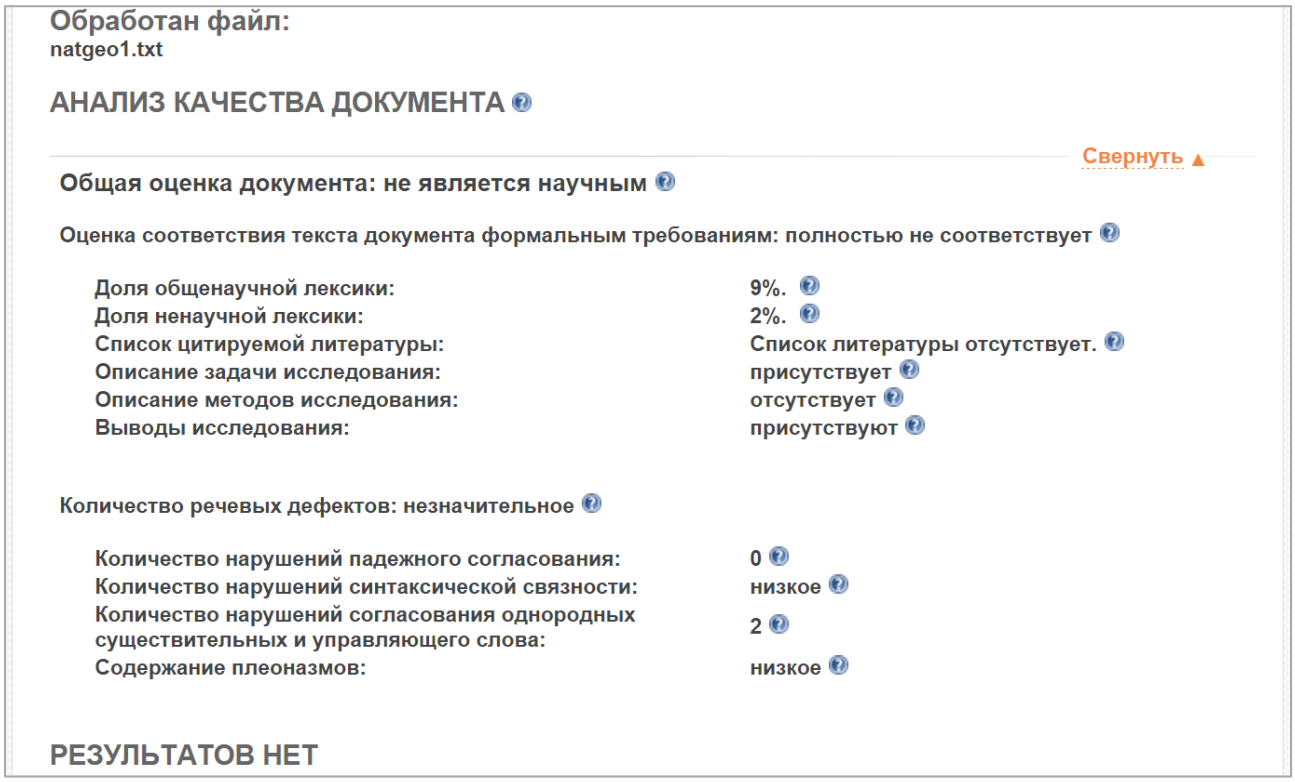


Рис.2 – Снимок отчета в системе Exactus Expert (http://expert.exactus.ru). Пример отчета для научно-популярного текста

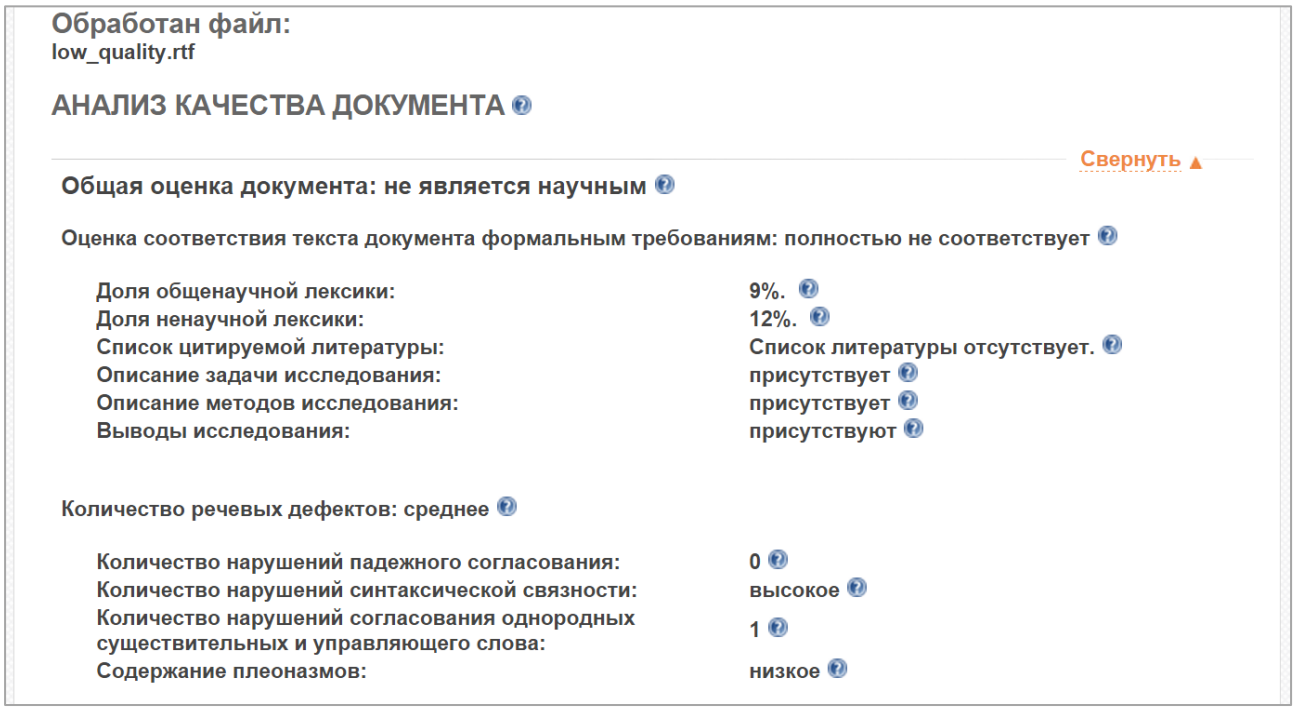


Рис.3 – Снимок отчета в системе Exactus Expert (http://expert.exactus.ru). Пример отчета для псевдонаучного текста