

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ СИСТЕМ ИНФОРМАТИКИ ИМ. А.П. ЕРШОВА
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи



Бакиева Айгерим Муратовна

**МОДЕЛИ ОПРЕДЕЛЕНИЯ ТЕМ ТЕКСТОВ,
ОСНОВАННЫЕ НА ГРАФАХ, И ИХ ПРИМЕНЕНИЕ
ДЛЯ РЕШЕНИЯ ЗАДАЧИ АВТОРЕФЕРИРОВАНИЯ**

Специальность 05.13.17 – Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
кандидат физико-математических наук, доцент
Батура Татьяна Викторовна

Новосибирск – 2018

Оглавление

Введение	4
Глава 1. Основные методы автореферирования	11
1.1 Экстрагирующие методы	17
1.2 Абстрагирующие методы	20
1.3 Гибридные методы	24
1.4 Выводы по главе 1	27
Глава 2. Основные понятия и постановка задачи построения тематических моделей	29
2.1 Построение модели текста	29
2.2 Построение тематической модели коллекции документов	31
2.3 Проблема согласования многословных терминов	35
2.4 Выводы по главе 2	39
Глава 3. Гибридный метод автоматического построения аннотаций научных текстов	40
3.1 Построение униграммных и расширенных тематических моделей	41
3.1.1 Выбор алгоритма тематического моделирования	41
3.1.2 Извлечение многословных терминов	43
3.1.3 Алгоритм построения расширенных тематических моделей	45
3.2 Риторический анализ и преобразования графов	47
3.2.1 Формальное описание преобразования текста	48
3.3 Операция сглаживания	52
3.4 Применение предложенных методов для обработки текстов на тюркских языках	54
3.4.1 Особенности морфологического анализа	55
3.4.2 Особенности синтаксического и риторического анализа	59
3.5 Выводы по главе 3	67
Глава 4. Оценка эффективности разработанных методов	69
4.1 Оценка тематических моделей и качества извлечения ключевых терминов	69
4.2 Оценка результатов реферирования	77
4.2.1 Метрика Rouge	77
4.2.2 Метрика RAV	78
4.2.3 Экспертная оценка	79
4.2.4 Точность, полнота, F-мера	79
4.3 Выводы по главе 4	80
Заключение	82
Список сокращений и условных обозначений	84

Литература	86
Приложения.....	97
Приложение А. Таблицы маркеров и коннекторов.....	97
Приложение Б. Шаблоны для сглаживания.....	108
Приложение В. Примеры работы системы	113
Приложение Г. Свидетельства о регистрации программ ЭВМ	117
Приложение Д. Акты о внедрении	120

Введение

Актуальность темы. Ввиду стремительного роста объемов текстовой информации, исследования в области компьютерной лингвистики на естественном языке сохраняют свою актуальность. На сегодняшний день наблюдается колоссальный рост количества информации, создаваемой людьми и машинами на естественном языке. Разработка алгоритмов и создание систем интеллектуального анализа данных, автоматического реферирования, поиска и извлечения информации, определения тем текстов, классификации и кластеризации текстовых документов по-прежнему являются сложными задачами.

Непрерывное увеличение интенсивности потока текстовой информации делает все более важной задачу семантического сжатия текстов. Связи между риторическими маркерами, коннекторами и ключевыми словами в тексте задают семантическую иерархию, которая позволяет решать различные задачи обработки текстов на естественном языке и является важным элементом при автореферировании и определении тем текстов.

В данной работе предложен гибридный метод автоматического построения аннотаций научных текстов в области информационных технологий, до сих пор остающейся за рамками внимания исследователей-разработчиков систем реферирования. Между тем реферирование статей по информационным технологиям особенно актуально, поскольку информационные технологии используются практически во всех отраслях науки и техники.

Таким образом, **актуальной** является задача создания новых методов автоматического построения аннотаций научных статей, решение которой служит приоритетным средством обмена информацией в процессе профессиональной коммуникации большого количества специалистов.

Степень проработанности темы. В настоящее время наблюдается большой научный интерес к области автоматизации реферирования и аннотирования. Этой проблемой начали заниматься во второй половине XX века такие ученые как H.P. Luhn, D. Marcu, K. Ono, U. Hahn, D. Radev, H. Saggion, L. Plaza, H.P. Edmundson, J. Kupiec, E. Lloret, J.J. Pollock, T. Strzalkowski, Р.Г. Пиотровский, В.П. Леонов, Д.Г. Лахути, Э.Ф. Скороходько, С.М. Приходько, В.А. Яцко, А.В. Анисимов, С.А. Тревгода, П.Г. Осминин, и др.

Среди российских исследователей наибольший вклад в данную область внесли научные группы, возглавляемые Н.В. Лукашевич, П.И. Браславским, С.О. Шереметьевой.

На сегодняшний день область научных исследований, связанная с автоматическим реферированием, продолжает активно развиваться.

Существует много путей решения этой задачи, которые довольно четко подразделяются на три направления: экстракция, абстракция и гибридный подход. Экстракция – извлечение из

исходного текста наиболее информативных предложений, т.е. формирование квазиреферата. Этот способ иногда называют поверхностным. Абстракция – обобщение текста первичного документа на достаточно высоком уровне посредством генерации текста реферата на основе абстрактного представления смысла; генерация текста реферата выполняется с учетом морфологии, синтаксиса, семантики, благодаря чему формируется логически и по смыслу связный текст. Этот способ называют глубинным. Гибридный подход сочетает в себе методы экстракции и абстракции.

Цель и задачи исследования. Целью данной работы является создание новых методов, применяемых для решения задачи автореферирования, описание формальных моделей и реализация основных компонентов системы для работы с научно-техническими текстами, ориентированной на генерацию корректного по содержанию текста реферата с правильной синтаксической структурой.

Поставленная цель достигается последовательным решением следующих **задач**.

1. Разработать метод формирования авторефератов на основе теории риторических структур.
2. Создать алгоритм построения расширенных моделей определения тем текстов.
3. Предложить метод извлечения наиболее значимых предложений из текста.
4. Описать процедуру сглаживания, позволяющую сделать текст полученной аннотации более связным и последовательным.
5. Реализовать разработанные модели, методы и алгоритмы в виде комплекса программ, позволяющего построить систему автореферирования на разных языках.
6. Провести вычислительные эксперименты, подтверждающие эффективность предложенных методов.

Соответствие диссертации паспорту специальности. Диссертация соответствует области исследований специальности 05.13.17 – Теоретические основы информатики по п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»; п. 6 «Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке»; п. 12 «Разработка математических, логических, семиотических и лингвистических моделей и методов взаимодействия информационных процессов, в том числе на базе специализированных вычислительных систем».

Методология и методы исследования. Методологической основой исследования является логика предикатов и теория графов. Для построения авторефератов и аннотаций использовались методы компьютерной лингвистики и машинного обучения. При разработке комплекса программ построения авторефератов и поиска определения тем текстов применялись методы машинного обучения и объектно-ориентированного программирования.

Научная новизна работы заключается в следующем:

- предложен гибридный метод построения аннотаций научных текстов, использующий представление текстов в виде графов;
- описана методика обнаружения важных элементов текста на основе теории риторических структур;
- предложен алгоритм построения расширенных моделей определения тем текстов на русском языке;
- создана лингвистическая база данных на основе анализа подязыка рефератов, используемая для определения весов предложений;
- описана процедура сглаживания, позволяющая сделать текст полученного реферата более связным и последовательным.

Актуальность и новизна исследования определяют его теоретическую и практическую значимость.

Теоретическая ценность работы состоит в том, что в ней дано формальное описание методов, алгоритмов и решений, позволяющих производить автоматическое построение лингвистических механизмов порождения нового текста строгой функциональной направленности на основе формального представления содержания.

Практическая значимость работы заключается в том, что на базе разработанных моделей создана система автоматического реферирования и аннотирования документов на разных языках. Разработанные методы, алгоритмы и программное обеспечение могут применяться для построения систем машинного понимания текста, систем автоматической обработки текста, информационно-поисковых систем и других информационных систем, основанных на знаниях.

Результаты диссертационной работы используются в исследованиях и разработках, проводимых в Лаборатории информационных ресурсов Института вычислительных технологий СО РАН, Лаборатории моделирования сложных систем Института систем информатики им. А.П. Ершова СО РАН и в компании «Новые программные системы», что подтверждается актами о внедрении.

1. Созданный программный комплекс используется для анализа больших наборов данных с целью автоматического извлечения важной информации по перспективным научным направлениям и технологиям. Данные представляют собой наборы до 60 тысяч файлов.

2. Отдельные связанные программные компоненты, разработанные А.М. Бакиевой в процессе работы над диссертацией, в частности, касающиеся машинного обучения, применяются в лабораториях при реализации других проектов.

Основные этапы исследования выполнены в рамках проектов и грантов: Грант Министерства образования и науки Республики Казахстан № 0115PK01422 «Разработка информационно-поискового тезауруса (с учетом морфологии казахского языка) в полнотекстовых базах данных по ИТ-технологиям»; Грант Министерства образования и науки Республики Казахстан № AP05133550 «Модели и методы семантического анализа и представления смысла текста в компьютерной лингвистике»; Интеграционный проект СО РАН № AAAA-A18-118022190008-8 «Модели и методы создания информационных систем, интегрирующих географическую и временную составляющие документов, согласованных с мировыми стандартами и тенденциями развития национальной и международной информационной инфраструктуры, интегрированных в открытое семантическое пространство».

Автором получено 3 свидетельства о регистрации программного для ЭВМ.

Положения, выносимые на защиту. На защиту выносятся следующие новые научные результаты:

1. Разработан гибридный метод, который позволяет получать рефераты (аннотации) высокого качества и определять темы текстов в виде набора ключевых терминов. Предложенный метод основан на использовании лингвистической базы знаний, графовом представлении текстов и машинном обучении.

2. Формально описана методика обнаружения важных элементов в тексте, базирующаяся на понятиях теории риторических структур. Создана лингвистическая база данных на основе анализа подязыка рефератов, используемая для оценки весов предложений квазиреферата.

3. Предложен алгоритм построения расширенных тематических моделей коллекций текстовых документов.

4. Описана процедура сглаживания предложений, позволяющая сделать текст полученного реферата (аннотации) более связным и последовательным.

5. Предложенные модели, методы и алгоритмы реализованы в виде системы, позволяющей автоматически формировать аннотации статей научно-технической тематики.

6. Собрана коллекция текстов научных статей на русском языке (около 1200 текстов) для проведения экспериментов. Проведены вычислительные эксперименты, подтверждающие высокую эффективность предложенных методов и алгоритмов.

Степень достоверности результатов. Все полученные результаты подтверждаются экспериментами, проведенными в соответствии с общепринятыми стандартами.

Апробация результатов исследования. Основные результаты работы были представлены на следующих международных, всероссийских и региональных научных конференциях: 10-я международная конференция по применению информационных и коммуникационных технологий (AICT-2016) (12-14 октября 2016 г. Баку, Азербайджан); 15-я международная научная конференция “Information Technologies and Management” (28-29 апреля 2017 г. Рига, Латвия); Международная конференция «Актуальные проблемы чистой и прикладной математики» (22-25 августа 2017, г. Алматы, Казахстан); 4-я Международная конференция по компьютерной обработке тюркских языков «TurkLang-2017» (18–21 октября, 2017, г. Казань); 2-я международная научная конференция «Информатика и прикладная математика» (26-29 сентября 2018, г. Алматы, Казахстан); 55-ая международная научная студенческая конференция (МНСК – 2017) (17-20 апреля 2017, г. Новосибирск); 54-ая международная научная студенческая конференция (МНСК – 2016) (16-20 апреля 2016, Новосибирск); 17-ая всероссийская конференция молодых учёных по математическому моделированию и информационным технологиям (УМ-2016) (30 октября – 3 ноября 2016, г. Новосибирск); 16-ая всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (УМ-2015) (28-30 октября 2015, г. Красноярск); Марчуковские научные чтения - 2017 (MSR 2017) (25 июня – 14 июля 2017, г. Новосибирск); Всероссийская научно-практическая конференция с международным участием «Интеллектуальный анализ сигналов, данных и знаний: методы и средства» (14-17 ноября 2017, г. Новосибирск); 18-ая всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (УМ-2017) (21-25 августа 2017, г. Иркутск); Всероссийская конференция «Big Data Conference» (13 сентября 2018, г. Москва); 16-ая российская конференция «Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017) (4-7 декабря 2017, г. Новосибирск).

Основные результаты диссертации докладывались и обсуждались на следующих научных семинарах: «Интеллектуальные системы» (ИСИ СО РАН), «Информационные технологии в задачах филологии и компьютерной лингвистики» (ИВТ СО РАН).

Публикации соискателя по теме диссертации. Основные результаты диссертации опубликованы более, чем в 30 научных работах, в том числе: 6 ВАК РФ, 2 Web of Science, 2 Scopus; докладывались автором на 14 научных конференциях (Рига, Баку, Алматы, Астана,

Москва, Иркутск, Новосибирск, Казань, Красноярск). Получено 3 свидетельства о государственной регистрации программ для ЭВМ. Основные результаты диссертации содержатся в работах [91-125].

Личный вклад. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Все представленные в диссертации результаты получены лично автором.

Объем и структура работы. Диссертация состоит из введения, четырех глав, заключения и пяти приложений. Полный объем диссертации составляет 122 страницы, включая 19 рисунков и 22 таблицы. Список литературы содержит 125 наименования.

Содержание работы. Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

В первой главе приведен обзор существующих методов автоматического реферирования, перечислены отечественные и зарубежные программные продукты, реализующие некоторые из методов. Рассмотрены две классификации. Согласно одной из них выделяют три направления: экстракция, абстракция, гибридный подход, согласно другой – все методы можно разделить на пять групп: статистические, алгебраические, графовые, когерентные, на основе машинного обучения. В данной главе проанализированы преимущества и недостатки каждой группы методов, сделаны выводы о целесообразности их использования.

Вторая глава посвящена построению моделей тем текстов и коллекций документов. Предложено решение проблемы многословных терминов с помощью графов, методов машинного обучения и других вспомогательных методов. Перечислены трудности, возникающие при построении тематических моделей.

Третья глава содержит описание предлагаемого гибридного метода автоматического построения аннотаций, моделей и методов, используемых в разработанной системе. В общем виде алгоритм состоит из следующих этапов: предварительная обработка текста; построение тематических моделей (униграммной и расширенной); риторический анализ и формирование квазиреферата; оценка весов предложений; выбор наиболее важных предложений; сглаживание полученного текста аннотации. Система реализована на языке Python3, также используется инструмент для работы с базами данных PostgreSQL. Используются внешние библиотеки Scikitlearn, Gensim, TensorFlow, NLTK, BigARTM, Flask и некоторые другие.

В данной главе также изложены методологические принципы применения предложенных методов для обработки текстов на тюркских языках, таких как казахский и

турецкий. Описаны особенности автоматизации морфологического и синтаксического анализа языков такого строя.

Четвертая глава посвящена проверке эффективности разработанных методов. Проверка эффективности осуществлялась посредством сравнения результатов автора с результатами, полученными путем использования методов, опубликованных в открытой литературе.

В заключении сделаны выводы и подведены итоги проведенного исследования.

В приложениях приведены таблицы дискурсивных маркеров и коннекторов, использованных в данной диссертационной работе; представлены шаблоны, применяемые для сглаживания текста аннотации; содержатся примеры результатов работы разработанной системы; представлены полученные свидетельства о регистрации и акты о внедрении.

Глава 1. Основные методы автореферирования

В настоящее время существует проблема информационной перегрузки. Автоматическое реферирование и аннотирование помогает человеку эффективно обрабатывать большие объемы информации. Рефераты и аннотации дают возможность установить основное содержание документа и определить необходимость обращения к первоисточнику. Поэтому в современном мире возрастает актуальность применения методов автоматического реферирования и аннотирования.

Автоматическое реферирование (Automatic Text Summarization) – извлечение наиболее важных сведений из одного или нескольких документов и составление их краткого описания. Алгоритм автореферирования – это преобразование, входными данными которого является текст (или несколько текстов), результатом является аннотация – сжатое представление этого текста. Вообще говоря, аннотация – краткая характеристика документа с точки зрения его назначения, содержания, вида, формы и других особенностей. Качество автоматической аннотации характеризуется разными параметрами: степень сжатия, логичность изложения, информативность, связность и др. Построение алгоритма автореферирования – наиболее трудная и вместе с тем нужная задача.

Существует много путей решения этой задачи, которые довольно четко подразделяются на три направления: экстракция, абстракция и гибридный подход. Экстракция – извлечение из исходного текста наиболее информативных предложений, т.е. формирование квазиреферата. Этот способ иногда называют поверхностным. Абстракция – генерация текста реферата с учетом морфологии, синтаксиса, семантики, благодаря чему формируется логически и по смыслу связный текст. Этот способ называют глубинным. Гибридный подход сочетает в себе методы экстракции и абстракции.

Глубинный способ формирования рефератов предполагает наличие методов синтаксического или семантического разбора предложений. В первом случае используются деревья синтаксического разбора. Процедуры автоматического реферирования манипулируют непосредственно деревьями, выполняя перегруппировку и сокращение ветвей на основании соответствующих критериев. Такое упрощение обеспечивает построение реферата – структурную выжимку исходного текста.

Во втором случае на этапе анализа также выполняется синтаксический разбор текста, но синтаксические деревья не порождаются, а формируются семантические структуры, которые накапливаются в виде концептуальных подграфов в базах знаний или тезаурусах. В частности, известны модели, позволяющие производить реферирование текстов на основе психологических ассоциаций сходства и контраста. В базах знаний избыточная и не имеющая

прямого отношения к тексту информация устраняется путем отсечения некоторых подграфов. Затем информация подвергается агрегированию методом слияния оставшихся графов или их обобщения. Для осуществления этих преобразований выполняются манипуляции логическими предположениями, выделяются определенные шаблоны в текстовой базе знаний. В результате преобразования формируется концептуальная структура текста в виде аннотации [1].

Многоуровневое структурирование текста с использованием семантических методов позволяет подходить к решению задачи реферирования различными путями.

1. Удаление малозначащих смысловых единиц. Преимуществом метода является гарантированное сохранение значащей информации, недостатком – низкая степень сжатия, т.е. сокращения объема реферата по сравнению с первичными документами.

2. Сокращение смысловых единиц – замена их основной лексической единицей, выражающей основной смысл.

3. Гибридный способ, заключающийся в уточнении реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Некоторые авторы [2] выделяют пять различных подходов к автореферированию (см. рисунок 1):

- статистический подход;
- когерентный подход;
- алгебраический подход;
- графовый подход;
- подход, основанный на машинном обучении.

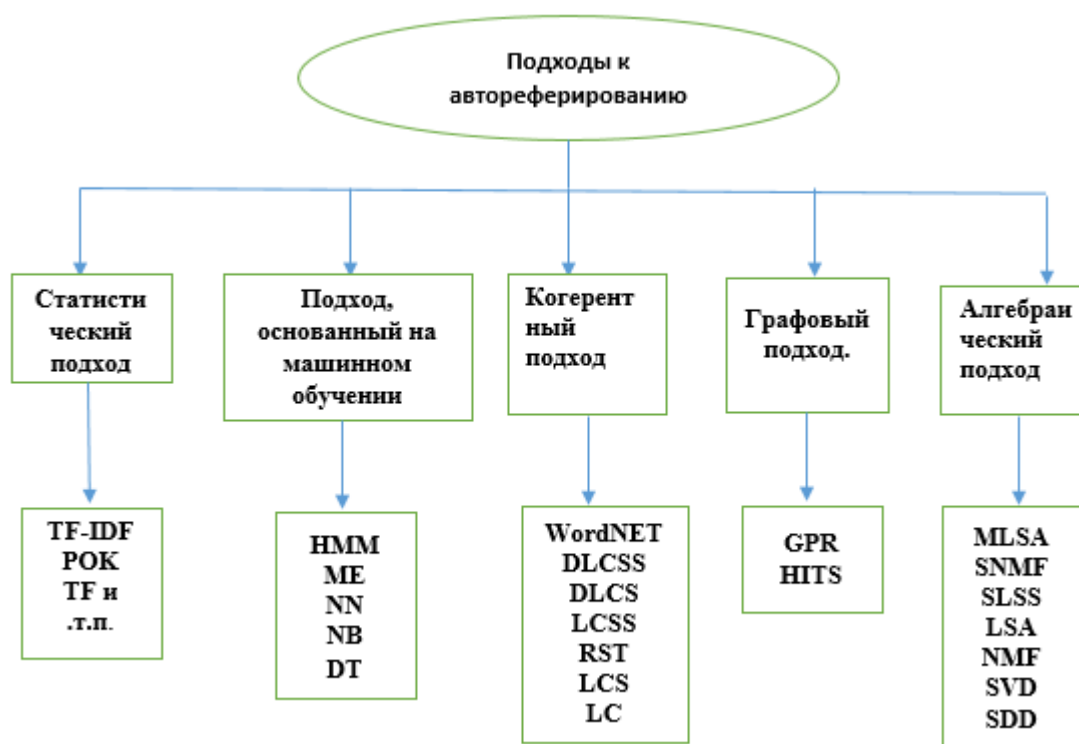


Рисунок 1 – Классификация подходов автореферирования текста

Статистический подход

Этот подход очень прост и часто используется для извлечения ключевых слов из документов. Для этого подхода нет предопределенного набора данных. Чтобы извлечь ключевые слова из документов, он использует несколько статистических характеристик документа, таких как частота слова (TF), временная частота обратных документов (TF-IDF), позиция ключевого слова (POK) и т. д.

Когерентный подход

Этот подход в основном касается отношений согласованности между словами. Сопряженные отношения между элементами в тексте: ссылка, эллипсис, замещение, союз и лексическая когерентность. Лексическая цепочка слова (LC), WordNet (WN), оценка лексической цепочки (LCS), оценка прямой лексической цепочки (DLCS), оценка диапазона лексической цепочки (LCSS), оценка диапазона прямой лексической цепочки (DLCSS), теория риторических структур (RST).

Алгебраический подход

В этом подходе используются алгебраические теории, а именно матрица, транспонирование матрицы, собственные векторы и т. д. Существует много алгоритмов, используемых для обобщения текста на основе алгебраического подхода, например, латентный семантический анализ (LSA), мета-латентный семантический анализ (MLSA), факторизация симметричных неотрицательных матриц (SNMF), семантический анализ уровня предложений

(SLSS), факторизация неотрицательных матриц (NMF), сингулярное разложение (SVD), полудискретное разложение (SDD).

Графовый подход

Графовый подход заключается в том, что фрагменты текста (слова, предложения, абзацы, в нашем случае – ЭДЕ) описываются в виде вершин графа, а отношения между вершинами (например, семантические отношения) обозначаются ребрами. Для обнаружения в тексте важных фрагментов, кроме того, используются такие популярные методы, основанные на графах, как: поиск гиперссылок с индуцированными темами (HITS) и Google PageRank (GPR).

Подход, основанный на машинном обучении

Машинное обучение – подход, характерной чертой которого является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для обучения нужен размеченный набор данных. Выходом алгоритма обучения является функция, аппроксимирующая неизвестную (восстанавливаемую) зависимость. Существует несколько популярных подходов к компьютерному обучению: метод Байеса (NB), деревья решений (DT), скрытая марковская модель (HMM), максимальная энтропия (ME), нейронные сети (NN), метод опорных векторов (SVM).

В такой классификации статистический и алгебраический подходы могут считаться экстракцией, когерентный и подход, основанный на машинном обучении, – абстракцией, а графовый подход является гибридным.

На международном рынке представлено множество программных продуктов, которые позволяют создавать авторефераты. Ориентированы они преимущественно на документы, содержащие текст на английском языке. В таблицах 1 и 2 приведены отечественные и зарубежные системы автоматического реферирования [3-9].

Таблица 1 – Отечественные системы автоматического реферирования и аннотирования, реализующие поверхностные методы

Наименования системы	Основные функции
ОРФО 8.0	Функция автоматического аннотирования русских текстов. Разработчик – компания «Информатик».
Либретто	Обеспечивает автоматическое реферирование и аннотирование русских и английских текстов; система встраивается в Word. Разработчик – компания

	«МедиаЛингва».
МедиаЛингва Аннотатор	Служит инструментарием для реализации функций автоматического реферирования и аннотирования в прикладных ИАС.
Следопыт	Поисковая система, включающая в себя средства автоматического реферирования и аннотирования документов.
Поисковая машина «Золотой Ключик»	Программная библиотека, работающая по принципу фильтрации на базе тезауруса. Как входные данные программе подается произвольный текст на русском языке, на стандартном выходе программа формирует аннотацию данного текста и список рубрик, к которым относится данный текст. В качестве аннотации используются предложения из входного текста, наиболее полно отражающие тематику текста. При рубрикации текста используется фиксированный список заранее определенных рубрик.
Inxight Summarizer	Выделяет наиболее весомые предложения из текста используя статистические алгоритмы, либо слова-подсказки.
eXtragon	Содержит набор исходных данных, созданный на основе оценивавшихся запросов для поиска по Веб-коллекции и по коллекции нормативно-правовых документов.
Galaktika-ZOOM	Интеллектуальный поиск по ключевым словам с учетом морфологии русского и английского языков, а также формирование информационных массивов по конкретным аспектам.
InfoStream	Технология позволяет создавать полнотекстовые базы данных и осуществлять поиск информации, формировать тематические информационные каналы, автоматически рубрицировать информацию, формировать дайджесты, таблицы взаимосвязей понятий (относительно встречаемости их в сетевых публикациях), гистограммы распределения весовых значений отдельных понятий, а также динамики их встречаемости по времени.
TextAnalyst	TextAnalyst работает только с русским языком, выделяя именные группы и строя на их основе семантическую сеть – структуру взаимозависимостей между именными группами. Программа создана в Московском научно-производственном инновационном центре «МикроСистемы».

Таблица 2 – Зарубежные системы автоматического реферирования и аннотирования

Наименование системы	Основные функции
Extractor	Использованы способы определения наиболее вероятных ключевых фраз; контекстная информация служит основой для идеи выявления в тексте переформулированных смысловых конструкций.
Autonomy Knowledge Server	Анализ текстов и идентификации ключевых концепций в пределах документов путем анализа корреляции частот и отношений терминов со смыслом текста.
InterMedia Text, Oracle Text	В ходе обработки текст каждого документа подвергается процедурам лингвистического и статистического анализа, в результате чего определяются его ключевые темы и строятся тематические резюме, а также общее резюме – реферат.
SemioMap	Поддерживает разбиение материала по папкам, создание отдельной базы данных для каждой папки. Связи между понятиями, которые выявляет SemioMap, базируются на совместной встречаемости фраз в абзацах исходного текстового массива.
Text Miner	Позволяет выбрать из потока информации необходимые данные и структурировать их. В качестве входных данных можно использовать не только текстовые документы или веб-страницы, но также ссылки, списки или кластеры.
WebAnalyst	Представляет собой интеллектуальное масштабируемое клиент/серверное решение для компаний, желающих максимизировать эффект анализа данных в Web-среде. Сервер WebAnalyst функционирует как экспертная система сбора информации и управления контентом Web-сайта. Модули WebAnalyst решают три задачи: сбор максимального количества информации о посетителях сайта и запрашиваемых ими ресурсах; исследование собранных данных и генерация персонализированного, на основе результатов исследований, контента.
Intelligent Text Miner (IBM)	Технология эффективного анализа текстовых данных. Представляет собой набор отдельных утилит, запускаемых из командной строки или скриптов независимо друг от друга. Данная система является одним из лучших инструментов глубинного анализа текстов.
Microsoft Word 97	Функция автоматического реферирования
Oracle Context	Разнообразие источников, форматов, запросов

RCO FX Ru	Программный продукт предназначен для аналитической обработки текста на русском языке. Основной сферой применения программы являются задачи из области компьютерной разведки, требующие высокоточного поиска информации. Например, к ним можно отнести автоматический подбор материала к досье на целевой объект или же мониторинг определенных сторон его активности, освещаемых в СМИ.
------------------	---

Перечисленные средства обеспечивают выбор фрагментов текста из исходных документов и соединение их в короткую аннотацию. Из рассмотренных программных продуктов на данный момент наименее гибким является «Золотой ключик». TextAnalyst как программный продукт, основанный на алгоритмах семантических сетей, проявляет значительно большую гибкость при работе с базами знаний и алгоритмами формирования смыслового портрета. Тот факт, что в «МедиаЛингва Аннотаторе» применяются алгоритмы определения семантических весовых коэффициентов предложений и специальные вероятностные модели, но при этом нет возможности создания смыслового портрета, позволяет относить МедиаЛингва Аннотатор [10] к промежуточному классу программных продуктов между «Золотой ключик» и TextAnalyst. Рассмотренная программа Extractor в большей степени подготовлена к работе в сети Интернет (например, в составе поисковых машин). Это делает Extractor более популярной и востребованной на международном рынке услуг автореферирования и поиска информации. Наибольшие перспективы в данной области видятся в развитии взаимодействия и совмещения алгоритмов формирования семантических сетей и алгоритмов поисковых машин в глобальной сети Интернет, и создание на базе совмещенных алгоритмов новых, общедоступных сервисов интеллектуального поиска информации, а также систем автореферирования больших объемов текстовой информации. Использование общедоступных сервисов по поиску и автореферированию позволит значительно облегчить задачу. Одним из возможных решений в этой ситуации может стать, создание систем составления краткого изложения полнотекстовых документов на базе общедоступных сервисов. Представляется возможным проектирование и разработка совмещённых поисковых систем с системами автореферирования.

1.1 Экстрагирующие методы

Тексты рефератов, как уже говорилось ранее, могут полностью состоять из предложений, извлеченных из исходного текста (полная экстракция), представлять собой комбинацию фрагментов исходного и нового текста, возможно даже в рамках одного предложения (сочетание экстракции и абстракции), а также не содержать предложений

исходного текста вообще (полная абстракция). В данной работе метод экстракции использовался для формирования квазиреферата.

В рамках квазиреферирования выделяют три основных направления, которые в современных системах применяются совместно:

- статистические методы, основанные на оценке информативности разных элементов текста по частоте появления, которая служит основным критерием информативности слов, предложений или фраз;
- позиционные методы, которые опираются на предположение о том, что информативность элемента текста зависит от его позиции в документе;
- индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний – маркеров важности, которые характеризуют их содержательную значимость. После выявления определенного (задаваемого, как правило, коэффициентом необходимого сжатия) количества текстовых блоков с наивысшими весовыми коэффициентами, они объединяются для построения квазиреферата [11].

Краткое изложение содержания первичных документов основывается на выделении из текстов наиболее важной информации и порождении новых текстов, содержательно обобщающих первичные документы. В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний, опирается на автоматизированный качественный контент – анализ, состоящий, как правило, из трех основных стадий. Первая – сведение исходной текстовой информации к заданному числу фрагментов – единиц значения, которыми являются категории, последовательности и темы. На второй стадии производится поиск регулярных связей между единицами значения, после чего начинается третья стадия – формирования выводов и обобщений. На этой стадии создается структурная аннотация, представляющая содержание текста в виде совокупности концептуально связанных смысловых единиц.

Экстрагирующие методы реферирования создают текст реферата на основе наиболее значимых текстовых фрагментов (предложения, абзацы) исходного документа. Значимость фрагментов может определяться по различным критериям, например, по содержанию во фрагменте ключевых слов, по расположению фрагмента в исходном тексте (заголовки, подзаголовки и т.д.), по наличию сигнальных фраз. При этом извлеченные фрагменты, во многих случаях, не обрабатываются, а извлекаются без изменений в порядке их следования в исходном документе.

Первая работа по автоматическому реферированию была сделана американским ученым Г. П. Луном в 1958 г. [12] на материале английского языка. Вес предложения, который Лун

называл «фактором важности» рассчитывался на основе двух показателей – частоты употребления слова и расстояния (количество слов) между ключевыми словами. Ключевыми словами считались наиболее частотные слова в тексте, за исключением стоп-слов.

В работах Н.В. Лукашевич [13-15] используется подход к реферированию на основе тезауруса. Для определенной области знания строится тезаурус, содержащий основные понятия этой области. Далее лексикон текста сравнивается с лексиконом тезауруса. На основе этого строится тематическое представление текста в виде тематических узлов – понятий, упоминаемых в тексте. Затем отбираются повествовательные предложения исходного текста, создается таблица всех возможных пар тематических узлов. В реферат из исходного текста извлекаются в порядке следования в тексте те предложения, в которых содержится пара не упоминавшихся ранее тематических узлов.

Таким образом, в этом подходе реферирование сводится к двум основным операциям: определению функционального веса (числа межфразовых связей) каждого предложения текста и отбору предложений, вес которых превышает некоторую пороговую величину. При определении числа межфразовых связей предложения используются четыре критерия семантической связи, все они сводятся к выявлению лексических и семантических повторов. С помощью первого критерия устанавливаются межфразовые связи на основе совпадения имен существительных. На основе второго (основного) критерия учитываются повторения существительных, производных от них имен прилагательных и глаголов, а также семантические связи типа «общее-частное» и языковые синонимы. Третий критерий расширяет второй критерий и учитывает местоимения и контекстуальные синонимы. Четвертый критерий устанавливает межфразовые связи на основе совпадения основ имен существительных или имен прилагательных, причастий, глаголов и наречий.

В работе [16] для получения краткой аннотации применяется симметричное реферирование – подход, при котором вес предложения вычисляется как функция от количества связей с другими предложениями. Под связью понимается наличие одного и того же ключевого слова в двух предложениях. При этом учитываются словоформы ключевых слов. Для симметричного реферирования необходим тематический словарь. В тексте исходного документа выявляются предложения, содержащие лексику словаря, затем подсчитываются сумма левосторонних и правосторонних связей и происходит извлечение предложений с наибольшим весом. В статье [17] отмечается, что симметричное реферирование применимо не только к научным текстам, но и к новостным текстам различного объема.

Г. Эдмундсон [18] в дополнение к критериям важности предложений Г. Луна добавил следующие: расположение предложения в документе и абзаце (заголовки, первые и последние предложения абзацев), присутствие сигнальных слов и выражений, таких как «важно»,

«определенно», «в частности», «неясно», «возможно», «например», присутствие слов из заголовка или подзаголовка.

Некоторые авторы [19, 20] рассматривают задачу реферирования как сокращение объема данных – исходный документ выступает в роли данных большой размерности, а задача реферирования – снизить размерность документа и сохранить основное содержание документа. Система Open Text Summarizer основана на похожих принципах [20]. Например, при использовании векторной модели для обработки текстовых данных получаются многомерные матрицы термины-на-документы и требуется их редукция. В качестве методов редукции используется латентно-семантический анализ, в основе которого лежит сингулярное разложение матрицы. Использование латентно-семантического анализа встречается в следующих работах [21-23].

К достоинствам экстрагирующих методов можно отнести независимость от предметной области, а также сравнительную простоту разработки: не требуется создания обширных баз знаний, проведения детального лингвистического анализа текста. К недостаткам экстрагирующих методов можно отнести то, что полученные рефераты часто являются бессвязными.

1.2 Абстрагирующие методы

Абстрагирующие методы анализируют исходный документ и опираются на лингвистическую базу знаний, на основе которой создается текст реферата. При использовании таких подходов текст реферата строится алгоритмом, основываясь на лингвистических правилах обработки языка и специфике нужной подобласти. Абстрагирующие методы могут сжать текст сильнее, чем экстрагирующие, но их разработка сложна: требуется технология генерации текста, основанная на лингвистических правилах обработки естественного языка. Абстрагирующие методы способны создавать новый текст, не представленный явно в тексте исходного документа на основе машинного обучения и когерентных подходов.

В 1990-х годах для задач автоматического реферирования начали применяться алгоритмы машинного обучения. Машинное обучение – это раздел искусственного интеллекта, направленный на создание алгоритмов, способных на основе некоторых признаков решить задачу новым, не заложенным в алгоритм способом. Преимущество использования машинного обучения заключается в удобстве тестирования целого ряда критериев оценки предложений. Первая работа в этом направлении была сделана в 1995 г. [24]. Авторы рассматривали задачу реферирования в качестве задачи классификации – включать или не включать предложения из текста статьи в реферат. В качестве критериев значимости предложений использовались: длина предложения, наличие имен собственных, расположение предложения в абзаце и другие.

Авторы сопоставили предложения рефератов и предложения статей при помощи разработанной программы. Полученный корпус использовался для обучения алгоритма. Алгоритм был основан на наивном баевсовском классификаторе, который маркировал каждое предложение: включить его в реферат или нет. Наивный байесовский классификатор – это простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости. Каждому предложению приписывался определенный вес, в соответствии со специальной формулой. В реферат входили n предложений с наибольшим весом.

В работе М. Кумара и др. [25] описывается система на основе машинного обучения, которая создает рефераты совещаний, исходя из текста и данных о событиях. Событиями служат записи в базе данных о назначении задания и завершении задания. Для генерации текста применяются шаблоны, для определения которых используются рефераты совещаний, написанные экспертами. После генерации всех шаблонов, чтобы выбрать их для включения в реферат, авторы использовали методы машинного обучения. Из 11 различных методов (Naive Bayes, Voted Perceptron, Support Vector Machines, Ranking Perceptron, K Nearest Neighbor, Decision Tree, AdaBoost, Passive Aggressive learner, Maximum Entropy learner, Balanced Winnow and Boosted Ranking learner) лучший результат показал метод Balanced Winnow.

В работе [26] исследуется проблема автоматической генерации структуры рефератов. Авторы отмечают, что предикаты и предикатные фразы имеют коммуникативную функцию – предупреждение читателя о содержании реферируемого документа путем явного указания («упоминает», «представляет», «предлагает»). Разработанный алгоритм получает на входе набор извлеченных фрагментов предложений и определяет, как соединить фрагменты в реферат. Из заранее определенного словаря на каждом шаге наиболее подходящий предикат (фраза) выбирается алгоритмом для вставки в начало текущего фрагмента. В работе используются различные алгоритмы машинного обучения (метод опорных векторов, наивный байесовский классификатор, деревья решений, метод ближайших соседей). Наилучшие результаты показал метод опорных векторов со следующим набором признаков для обучения: позиционные признаки (расположение вставляемого предиката в реферат), количество слов в предложении, присутствие в предложении слов из заголовка, содержательные признаки (синтаксически главный элемент именной или глагольной фразы). Оценка результатов показала, что разработанный алгоритм может прогнозировать структуру рефератов более чем в 60 % случаев.

Работа [27] выполнена на материале арабского языка и основана на сочетании машинного обучения, статистического анализа и анализа риторических структур. Сначала выполняется риторический анализ и определение единиц для извлечения, затем осуществляется

классификация методом опорных векторов (SVM), чтобы выбрать, какие единицы перенести в реферат.

Как уже отмечалось ранее, к когерентному способу относится использование теории риторических структур (TPC). В TPC [28] в качестве семантических отношений рассматриваются риторические отношения. Данная теория основана на предположении о том, что любая единица дискурса связана с другой единицей данного дискурса посредством некоторой осмысленной связи. Таким образом, основными понятиями TPC являются дискурсивная единица и отношение. В TPC определено два типа ЭДЕ: ядро и сателлит. Ядро рассматривается в качестве наиболее важной части высказывания, тогда как сателлиты поясняют ядра и являются вторичными. Ядро содержит основную информацию, а сателлит содержит дополнительную информацию о ядре. Сателлит часто бывает непонятным без ядра. В то время как выражения, где сателлиты были удалены могут быть поняты в определенной степени. Последовательные ЭДЕ соединяются между собой риторическими отношениями. Эти части являются элементами, из которых строятся более крупные фрагменты текстов и целые тексты. Каждый фрагмент по отношению к другим фрагментам выполняет определенную роль. Текстовая связность формируется посредством тех отношений, которые моделируются между фрагментами внутри текста.

Согласно данной теории любой текст может быть представлен в виде графа, узлами которого являются элементарные дискурсивные единицы (ЭДЕ – a unit) или группы таких единиц – дискурсивные единицы (ДЕ – a text span). При этом вне зависимости от уровня иерархии, узлы графа будут связаны одним и тем же набором отношений на уровне выше отдельного предложения. Такие связи называются риторическими отношениями.

В работе 1994 г. [29] предлагается вычислительная модель дискурса для японских информативных текстов. В данной работе предлагается практическая процедура извлечения риторической структуры дискурса. Риторическая структура представляется в виде дерева. Процессу составления реферата предшествует извлечение риторических структур из текста статьи и их анализ. Оценка результатов показала, что в получаемых рефератах содержится до 74% важных предложений оригинальной статьи.

Д. Марку [30] в 1998 г. предложил оригинальный подход, основанный на теории риторических структур, для определения важных элементов в тексте. В работе использовались эвристические правила, основанные на дискурсе, наряду с традиционными признаками, которые используются для автоматического реферирования. Автор представляет входной текст в виде набора деревьев и предлагает использовать алгоритм ограничений для объединения этих деревьев. Далее применяется несколько эвристик для выбора более подходящих деревьев при формировании реферата. Автор отмечает, что различие между ядром и сателлитом основано на

эмпирическом наблюдении того факта, что ядро выражает более важную часть текста, чем сателлит. Также, ядро не зависит от сателлита, но не наоборот. Марку описывает риторический парсер, который строит дискурсное дерево. После создания дерева, можно получить частичное представление о расположении важных частей текста. Если задано условие, что реферат должен содержать k % текста, то первые k % частей из частичного представления может быть отобрано для реферата. В работе [31] авторы объединяют теорию дискурса и традиционные методы автоматического реферирования.

Попытки применения дискурсивного анализа для решения различных задач компьютерной лингвистики можно заметить в современной практике. Подробный обзор литературы, представленной в статье [32], показывает, что в большинстве случаев дискурсивный анализ способен повысить качество автоматических систем на 4-44% в зависимости от конкретной задачи.

Система автореферирования научных статей, основанная на дискурсивном анализе, описана в [33]. В ней определены семь риторических категорий. Автор работы [34] применил теорию риторических структур для создания графического представления документа. На основе структурного анализа текста вычисляются веса предложений, из которых в итоге получается краткая аннотация. В работе [35] обсуждается создание реферата, содержащего информацию не только из одного конкретного документа, но и дополнительные знания из других, похожих на него по тематике документов.

Митхун С. описывает подход, базирующийся на схемах, для формирования аннотаций на основе запросов, в которых используются структуры дискурса [36]. Этот подход выполняет четыре основных задачи, а именно: категоризацию вопроса, идентификацию риторических предикатов, выбор схемы и обобщение. Автор создал систему под названием BlogSum и оценил ее производительность относительно релевантности и согласованности вопросов. Полученные результаты показывают, что предлагаемый подход решает проблему несоответствия и дискурсивной несогласованности автоматически созданных рефератов.

Исследования в этой области для английского языка достигли достаточно высокого уровня, но для текстов на русском языке данная область изучена сравнительно мало. Анализ подходов для решения проблемы автоматического формирования рефератов научно-технических текстов на русском языке проводился российскими учеными в работах [37] и [38]. В исследовании [37] описаны методы и алгоритмы, учитывающие нелинейный и иерархический характер текста. С помощью риторических отношений решается проблема экстракции (извлечения фрагментов текста). Тревгода С.А. разработал систему, основанную на правилах вывода и узкоспециализированном словаре ключевых фраз. Гибридный подход, предложенный Осмининым П.Г. [38], сочетает методы экстракции и абстракции. Этот подход был реализован

автором в системе реферирования, ориентированной на автоматический перевод. Описанная система построена для текстов по теме «математическое моделирование». Были использованы не только риторические структуры, но и глаголы из предметной области «математической логики». С помощью найденных ключевых слов определяется вес предложения, затем полученная аннотация формируется в соответствии с шаблонами.

Некоторые особенности риторических отношений описаны в работах [93, 109]. Там также формулируются утверждения о их свойствах. Работа [39] описывает опыт построения корпуса на русском языке, содержащего дискурсивные маркеры. Корпус общедоступный и включает в себя тексты разных жанров, таких как научный, научно-популярный и новостной. Прежде чем использовать теорию риторических структур, приходится адаптировать ее для конкретного языка. Это связано с грамматическими особенностями. В своей статье авторы предлагают иерархию риторических отношений, которая, согласно их исследованиям, является наиболее удобной и корректной для работы с текстами на русском языке.

В настоящее время абстрагирующие методы активно развиваются. В работах [40-42] также предлагаются методы автореферирования на основе абстрактного представления текста. Преимущества абстрагирующих методов заключаются в получении реферата более высокого качества, чем при применении экстрагирующих методов. К недостаткам данных методов относится сложность их практической реализации, необходимость сбора большого количества лингвистических знаний.

1.3 Гибридные методы

В наше время с целью улучшения работы над недостатками экстрагирующих и абстрагирующих методов разрабатываются гибридные методы автоматического реферирования. В гибридных методах извлеченные из первоисточника предложения (или их части) обрабатываются определенным образом, например, некоторые части предложений опускаются, выполняется слияние предложений, предложения переносятся в реферат в порядке, отличном от порядка следования в первоисточнике и т. д. Например, в системе COMPENDIUM [43] гибридный подход реализуется следующим образом: на вход подается реферат, составленный по экстрагирующей методике. Для этого реферата строится взвешенный граф, вершины которого представлены словами, а дуги отражают отношение смежности между словами. Вес дуг определяется по алгоритму PageRank. Затем между вершинами графа строится кратчайший путь с помощью алгоритма Дейкстры, таким образом, создается набор предложений-кандидатов. Следующий этап заключается в фильтрации неправильных путей. Авторы выделили следующие критерии правильных предложений: длина предложения не менее трех слов, в каждом предложении должен быть глагол, предложение не должно

оканчиваться на артикль, предлог, местоимение или союз. На последнем этапе происходит выбор предложений для включения в новый реферат из реферата, составленного по экстрагирующей методике или из набора предложений-кандидатов.

Наглядным примером гибридного способа построения системы автореферирования является многоязычная система SUMMARIST, описанная в [44]. Эта система сочетает в себе методы понятийного уровня знаний о мире, методы информационного поиска и статистические методы. Алгоритм состоит из трех этапов: идентификация темы, интерпретация и генерация. SUMMARIST формирует аннотации на пяти языках: английском, японском, испанском, индонезийском и арабском.

Также существует гибридная система SumUM [45], которая генерирует рефераты для научно-технических документов. Авторы провели исследование корпуса рефератов, выполненных людьми, и выявили ряд трансформаций, которые применяли референты, например, слияние информации из различных частей документа, перефразирование оригинала.

Подход авторов [46] к реферированию основывается на поверхностном анализе исходного документа, извлечении информации определенного вида и выполнении генерации текста. В системе также используются: маркировщик частей речи – лингвистические и концептуальные шаблоны, заданные регулярными выражениями; синтаксические категории; концептуальный словарь.

В работе [47] предложен метод реферирования, основанный на преобразовании текста в концепты с последующим представлением документа в виде графа. Метод использует дополнительные ресурсы – тезаурус медико-биологической области UMLS [48] и программу MetaMap [49] для преобразования текста в концепты из тезауруса UMLS. Метод состоит из следующих шагов: представление документа в виде графа, кластеризация концептов, выбор предложений. В первую очередь документ представляется в виде графа, где узлы являются концептами тезауруса UMLS, а ребра обозначают отношения между узлами. Для этого все предложения документа обрабатываются программой MetaMap, концепты UMLS дополняются своими гиперонимами. Далее каждому узлу присваивается оценка прямо пропорциональная глубине иерархии концептов. После этого все графы предложений объединяются в один граф документа. Затем выполняется кластеризация концептов. Каждый кластер представляет собой набор близких по значению концептов и может рассматриваться как тема документа. Процедура выбора предложений основывается на сходстве между кластерами и предложениями. Для выбора предложений авторы используют несколько эвристик.

Естественный язык очень сложен для автоматической обработки, поэтому исследователи, как правило, стремятся для улучшения качества получаемых результатов решать задачи реферирования для определенных предметных областей.

Авторы работы [50] исследуют задачу реферирования для текстов судебных решений. На основе анализа 3500 судебных решений на английском и французском языках и их рефератов, составленных профессиональными референтами, авторы выявили, что типичное судебное решение состоит из следующих разделов: данные о решении (имена, реквизиты сторон), вводная часть (информация о событиях, действиях лиц), основное содержание (изложение фактов в хронологическом порядке), правовой анализ (комментарии судьи), заключение (окончательное решение суда). Предлагаемая авторами система реферирования выполняет следующие действия: тематическое разбиение текста решения на основе лексических маркеров каждого раздела, фильтрацию материала – система пропускает фрагменты текста, не содержащие релевантную для реферата информацию (авторы указывают, что примерно 30 % документа не содержит релевантную для реферата информацию); отбор предложений – система оценивает релевантность предложений на основе простых эвристических правил (положение параграфа в тексте документа, расположение предложения в параграфе, мера TF-IDF и др.), генерация текста реферата – извлечение предложений и их представление в табличном формате. Длина реферата примерно составляет 10 % от объема исходного документа. По оценке экспертов, система правильно выявила 70 % предложений или параграфов. Реферированию юридических текстов посвящены также работы [51, 52].

Авторы работы [53] предлагают подход к реферированию оценочных суждений или комментариев пользователей Интернета. Авторы собрали корпус оценочных комментариев пользователей из отзывов на сайтах Amazon.com, WhatCar.com и социальной сети Twitter. Авторы работали с английским языком, тексты отзывов были посвящены сотовым телефонам и автомобилям. Собранный корпус был вручную размечен экспертом, который определял тональность комментария (отрицательный, нейтральный, положительный комментарий) и интенсивность оценки.

Авторы работы [54] предлагают гибридный подход к реферированию текстов патентов на английском, французском и немецком языках. Система сначала отбирает из текста патента предложения, а затем выполняет их слияние в связный текст, причем при слиянии учитывается не только лингвистическая информация, но информация о структуре документа (самостоятельный или зависимый пункт формулы изобретения и т.д.). Реферированию патентов посвящены также работы [55-59].

К гибриднему подходу могут быть отнесены графовые методы, когда фрагменты текста (слова, предложения, абзацы, в нашем случае ЭДЕ) описываются в виде вершин графа, а отношения между вершинами (например, семантические отношения) обозначаются ребрами. Примерами работ в этом направлении могут служить [60, 61].

Сложность при разработке гибридных методов заключается в выборе наиболее удачного сочетания методик генерации и извлечения. Гибридные методы по сравнению с абстрагирующими методами проще в разработке, а по сравнению с чисто экстрагирующими методами могут обеспечить лучшее качество выходного результата.

1.4 Выводы по главе 1

В данной главе рассмотрены основные методы автоматического реферирования, приведены две классификации существующих подходов, перечислены отечественные и зарубежные программные продукты, реализующие некоторые из методов.

В большинстве работ выделяют три основных подхода к автоматическому реферированию:

- экстрагирующие методы, основанные на извлечении из первичных документов наиболее информативных фрагментов и включении их в реферат в порядке следования в тексте;
- абстрагирующие методы, обобщающие текст первичного документа на достаточно высоком уровне посредством генерации текста реферата на основе абстрактного представления смысла; для генерации текста используются знания о морфологии, синтаксисе и семантике конкретного языка;
- гибридные методы, которые сочетают экстракцию и элементы абстракции.

Первые методы автоматического реферирования были экстрагирующими, то есть формировали текст реферата на основе наиболее значимых текстовых фрагментов исходного документа. К достоинствам экстрагирующих методов можно отнести независимость от предметной области, а также сравнительную простоту разработки: не требуется создания обширных баз знаний, проведения детального лингвистического анализа текста. К недостаткам экстрагирующих методов можно отнести то, что полученные рефераты часто являются бессвязными.

Абстрагирующие методы начали развиваться позднее и остаются на уровне исследовательских разработок, что связано со сложностью создания и ограниченностью предметной области. Абстрагирующие методы помогают создавать новый текст, не представленный явно в исходном документе, следовательно, повышается степень сжатия исходного документа. Преимущества абстрагирующих методов заключаются в получении реферата более высокого качества, чем при применении экстрагирующих методов. К недостаткам данных методов относится сложность их практической реализации, необходимость сбора большого количества лингвистических знаний.

С целью преодоления недостатков абстрагирующих и экстрагирующих методов разрабатываются гибридные методы автоматического реферирования, которые сочетают в себе стороны вышеуказанных подходов. Например, сначала происходит извлечение наиболее значимых фрагментов и их последующая обработка, потом осуществляется слияние предложений, удаление неинформативных частей и т.д. Сложность при разработке гибридных методов заключается в выборе наиболее удачного сочетания методик генерации и извлечения. Гибридные методы по сравнению с абстрагирующими методами проще в разработке, а по сравнению с чисто экстрагирующими методами могут обеспечить лучшее качество выходного результата.

Следует заметить, что так как естественный язык очень сложен для автоматической обработки, то при решении задачи автореферирования для улучшения качества получаемых результатов исследователи стараются ориентироваться на определенную предметную область.

Глава 2. Основные понятия и постановка задачи построения тематических моделей

2.1 Построение модели текста

Модель «граф слов». Предположим, что произвольный текст представляет собой последовательность слов, связанных друг с другом разными отношениями, например, синтаксическими, семантическими, ассоциативными и др. Находясь на уровне синтаксического анализа, не удастся в полной мере добиться решения многих задач автоматической обработки текста, таких как машинный перевод, автоматическое реферирование и пр. Необходимо использовать дополнительную информацию, например, знания о внешнем мире или рассматривать другие отношения между словами (помимо синтаксических). Примером таких отношений являются риторические. Их можно рассматривать как разновидность семантических отношений. О них будет рассказано далее.

Информация о порядке следования слов в тексте вместе с информацией о значимости слов также задает своего рода отношения между словами. Придание веса каждому слову в зависимости от его важности позволяет упорядочить слова в тексте от более значимых (которые передают основной смысл текста) к менее значимым (являются общими или вспомогательными при передаче смысла). Естественно, что более значимые, т.е. ключевые слова, в большей степени отражают смысл текста, его тематическую принадлежность. Поэтому поиск ключевых слов является неотъемлемым этапом в определении тем текстов.

Итак, текст можно представить в виде графа, где вершины – это фрагменты текста (отдельные слова, многословные выражения, дискурсивные единицы), а ребра отражают отношения между этими фрагментами. Каждая вершина имеет вес в зависимости от значимости фрагмента, чем больший вес, тем более значимый фрагмент. Отдельного пояснения требуют случаи многословных выражений. Под многословными выражениями понимаются последовательности двух или более лексем (слов), которые обладают свойствами отдельных лексем. В научных текстах это так называемые ключевые фразы или многословные термины. Примером являются словосочетания: «алгебраическое поле», «система уравнений», «компьютерная лингвистика» и др. Другой наиболее подходящий англоязычный термин – коллокация – это слова, которые обычно используются друг с другом, формируя устойчивое словосочетание. Далее в работе рассмотрены методы обнаружения ключевых слов и многословных терминов. Продолжим пока рассуждения о представлении текстов в виде графов.

Рассмотрим следующий пример.

Название научной статьи: «Программная конвейеризация циклов для ускорителя плавающей арифметики в составе процессора КОМДИВ128-РИО».

Фрагмент аннотации: «Использование точного подхода, основанного на применении методов целочисленного линейного программирования, позволяет обеспечить оптимальную производительность кода. Основное внимание в статье уделено особенностям формулировки задачи целочисленного линейного программирования, связанным со спецификой архитектуры ускорителя. Рассмотрены вопросы точного подсчета числа требуемых регистров, а также проблема понижения кратности развертки конвейеризованных циклов».

Ключевые слова, указанные авторами: «оптимизация программ, конвейеризованные циклы, целочисленное линейное программирование, архитектура ускорителя».

На рисунке 2 приведена графовая модель текста для определения его темы.

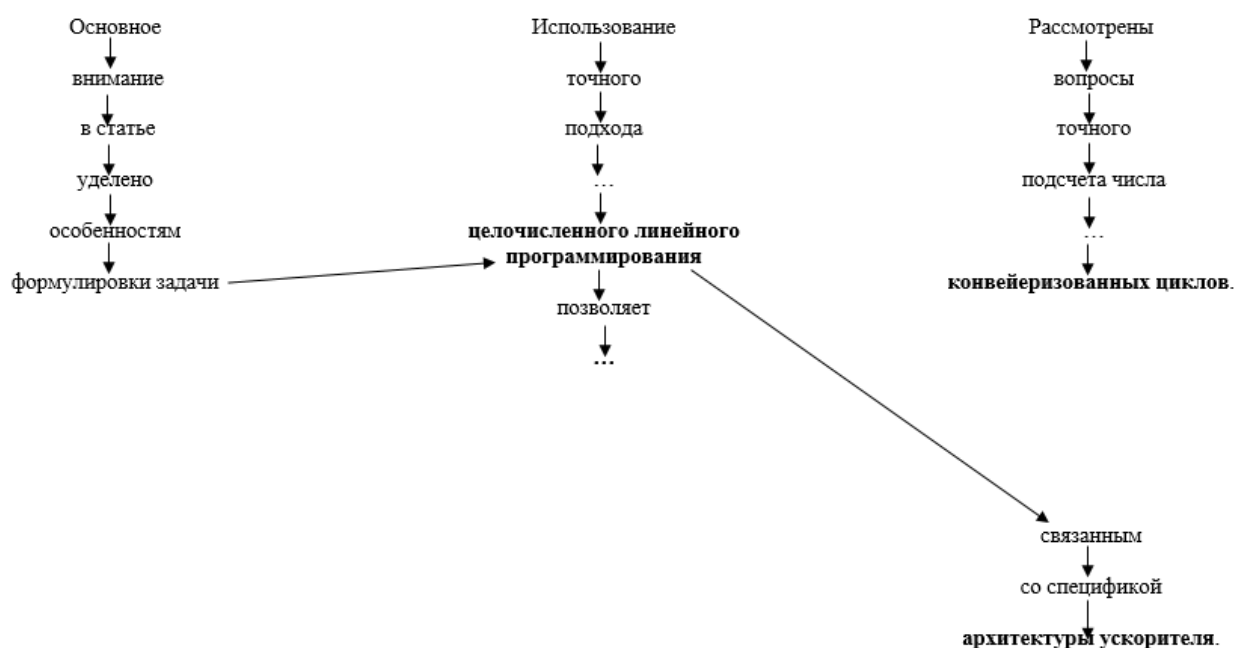


Рисунок 2 – Графовая модель текста

Видно, что ключевые термины, имеющие наибольший вес и выделенные жирным шрифтом на рисунке, описывают тему рассмотренной статьи.

Подобная идея положена в основу методов тематического моделирования. Эти методы подробно рассмотрены в следующих разделах данной работы.

Модель «граф тем». Такая модель основана на представлении заранее заданных тематических классов, образующих строго определенную иерархию: от более общей темы к более конкретной. Примером такой графовой модели является классификатор УДК, который широко используется для систематизации и поиска нужных сведений по конкретным темам и для группировки статей, публикаций и книг по тематическим разделам (см. рисунок 3).

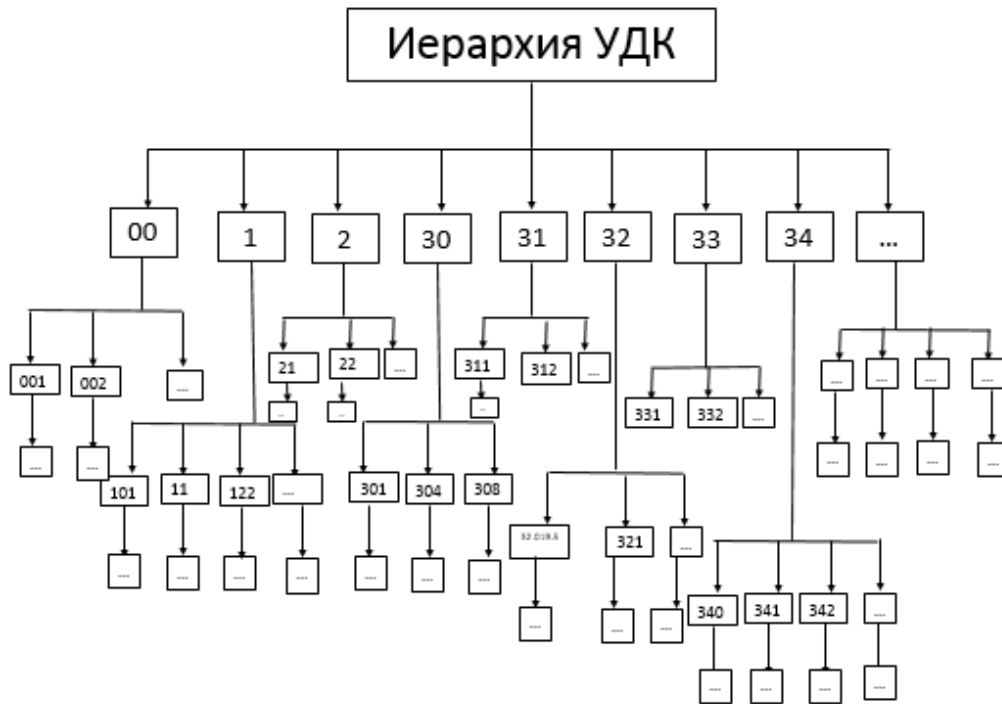


Рисунок 3 – Фрагмент иерархии тем

При наличии заданной иерархии тем определение темы текста сводится к задаче классификации, для решения которой могут применяться такие известные методы как метод опорных векторов (SVM), байесовский классификатор (NB), метод ближайших соседей (k-NN), деревья решений (DT) и др. Эти методы многократно показали свою эффективность, кроме того, они позволяют ранжировать тексты коллекции по критерию наилучшего попадания в класс от наиболее близкой темы к наиболее далекой.

Главное отличие между описанными выше моделями определения тем текстов состоит в том, что в первом случае отсутствует какая-либо априорная информация о темах, темы никак не заданы, не охарактеризованы, не имеют названий, их количество неизвестно. Во втором же случае предполагается наличие вполне конкретного набора тем, каждая из которых имеет название, и все они выстраиваются в определенную иерархию. Наличие априорной информации о темах в этом случае является обязательным. Все дальнейшие рассуждения выполнены в рамках первой модели.

2.2 Построение тематической модели коллекции документов

Тематическое моделирование – построение тематической модели некоторой коллекции текстовых документов. Тематическая модель представляет собой описание коллекции с помощью тематик, использующихся в документах этой коллекции, и определяет нужные термины, относящиеся к каждой из тематик [62]. В такой модели каждая тема представляется

дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем.

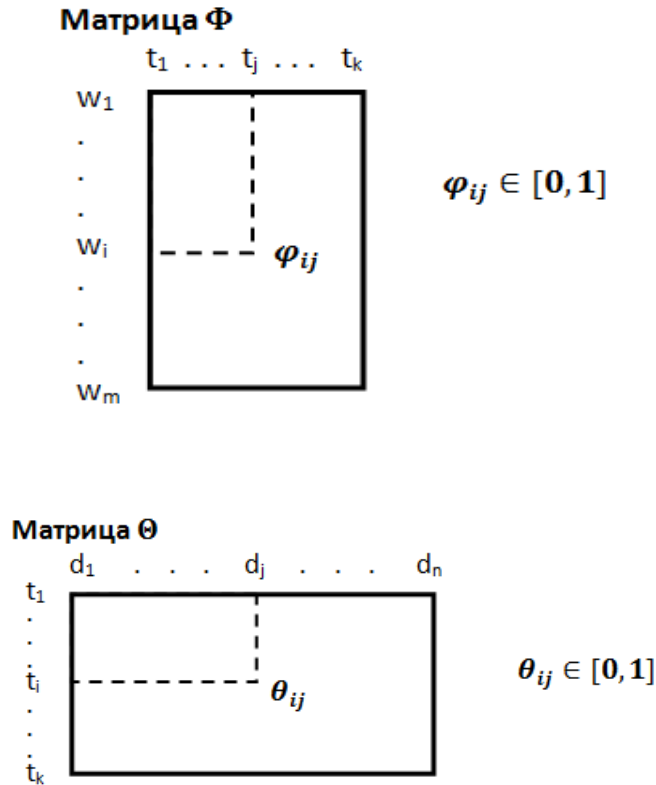
Вероятностная тематическая модель представляет каждую тему как дискретное распределение на множестве слов, а документ – как дискретное распределение на множестве тем [63].

Тема – набор терминов (слов и словосочетаний), характеризующих принадлежность текста к определенной области знаний. Эти термины правильнее называть тематическими словами. С нашей точки зрения, их функция несколько отличается от общепринятых ключевых слов. Тематические слова обязательно присутствуют в тексте и характеризуются частотой встречаемости. В отличие от них ключевые слова, заданные авторами, могут вообще отсутствовать в тексте. Однако, чтобы не отходить от общепринятой терминологии в дальнейших рассуждениях условимся называть тематические слова ключевыми.

Под многословным термином в данной работе подразумевается устойчивая последовательность слов (n -грамма), имеющая определенную семантику в контексте заданной предметной области, относящаяся к одной из выявленных в тексте тем и обладающая значительной частотой встречаемости по сравнению с другими n -граммами.

Пусть задана некоторая коллекция документов D , тогда W – множество всех встречающихся в данной коллекции терминов (слов или n -грамм). Каждый документ $d \in D$ представляется в виде последовательности терминов (w_1, \dots, w_{n_d}) длиной n_d , $w \in W$, при этом каждое ключевое слово может встретиться в документе несколько раз.

Предполагается, что существует некоторое множество тем T , причем каждое вхождение термина w связано с некоторой темой t . Коллекция документов рассматривается как множество троек (d, w, t) , выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. При этом документы $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, а тема $t \in T$ является скрытой переменной (см. рисунок 4).

Рисунок 4 – Матрицы Φ и Θ

Гипотеза «мешка слов» состоит в том, что элементы выборки независимы, т.е. порядок слов в тексте документа не имеет значения, а значит, тематическую модель можно выявить даже при произвольной перестановке терминов в тексте. В этом случае каждый документ представляется как подмножество $d \subseteq W$, в котором каждому элементу w_d поставлено в соответствие количество вхождений n_{dw} термина w в документ d .

Согласно определению условной вероятности, формуле полной вероятности и гипотезе условной независимости:

$$p(w | d) = \sum_{t \in T} p(w | t) \cdot p(t | d).$$

Тогда задача построения тематической коллекции документов заключается в том, чтобы найти для известной коллекции D множество всех использующихся в ней тем T , а также для каждого $d \in D$ по распределению слов по документам $p(w|d)$ восстановить распределения тем в документе $p(t|d)$ и слов по темам $p(w|t)$.

Тематические модели позволяют автоматически систематизировать большие коллекции текстовых документов на естественном языке, повышают эффективность информационного поиска. В данное время тематические модели находят применение в самых различных областях [64]: для создания персонализированных медицинских рекомендаций, для библиографического анализа, для анализа данных из социальных сетей, для многоязычного информационного

поиска, для выявления трендов в новостных потоках или научных публикациях, для автоматического присвоения тегов веб-страницам, в рекомендательных системах, учитывающих контекст, в анализе террористической активности в сети Интернет и многих других исследованиях.

Среди подходов к тематическому моделированию основными на сегодняшний день являются: PLSA (Probabilistic Latent Semantic Analysis, вероятностный латентный семантический анализ), LDA (Latent Dirichlet Allocation, латентное размещение Дирихле) и ARTM (Additive Regularization for Topic Modeling, аддитивная регуляризация тематических моделей).

PLSA – вероятностная тематическая модель представления текста на естественном языке. Модель называется латентной, так как предполагает введение скрытого (латентного) параметра, являющегося темой. Впервые описана в Томасом Хофманном в 1999 году [65].

LDA – модель, позволяющая объяснять результаты наблюдений с помощью неявных групп, благодаря чему возможно выявление причин сходства некоторых частей данных. Например, если наблюдениями являются слова, собранные в документы, утверждается, что каждый документ представляет собой смесь небольшого количества тем, и что появление каждого слова связано с одной из тем документа [66].

ARTM – аддитивная регуляризация тематических моделей, является обобщением большого числа алгоритмов тематического моделирования. ARTM позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. При таком подходе PLSA представляет собой тематическую модель без регуляризаторов, а LDA – тематическую модель, в которой каждая тема сглажена одним и тем же регуляризатором Дирихле. Предложена модель ARTM в 2014 году [67]. В настоящее время ARTM приобретает все большую популярность благодаря своей универсальности и гибкости настройки параметров моделей.

Современные требования к тематическим моделям довольно разнообразны. Основное из них заключается в том, что тематические модели должны хорошо поддаваться интерпретации, конечному пользователю должно быть понятны причины выделения определенных тем в тексте и структура самих тем. Эта особенность является главным преимуществом тематических моделей перед набирающими популярность нейронными сетями. Кроме того, часто требуется, чтобы тематические модели учитывали разнородные данные, выявляли динамику тем во времени, автоматически разделяли темы на подтемы, использовали не только отдельные ключевые слова, но и многословные термины и т.д.

Приписывание веса каждому слову в зависимости от его важности позволяет упорядочить слова в тексте от более значимых (которые передают основной смысл текста) к

менее значимым (являются общими или вспомогательными при передаче смысла). Естественно, что более значимые, т.е. ключевые слова, в большей степени отражают смысл текста, его тематическую принадлежность. Например, по набору ключевых слов, как правило, можно определить, к какой области относится описанное в статье исследование. Поэтому поиск ключевых слов является неотъемлемым этапом в определении тем текстов.

Отдельного пояснения требуют случаи многословных выражений. Под многословными выражениями понимаются последовательности двух или более лексем (слов), которые обладают свойствами отдельных лексем. В научных текстах это так называемые ключевые фразы или многословные термины. Примерами являются словосочетания «задача оптимизации», «извлечение знаний», «онтологическое моделирование» и т.д. Обнаружение их в тексте является отдельной подзадачей, о решении которой пойдет речь далее.

2.3 Проблема согласования многословных терминов

Как было упомянуто выше, основным требованием к тематическим моделям является их интерпретируемость. При этом в большинстве алгоритмов тематического моделирования в качестве терминов используются только слова, а не n-граммы. Также для человека использование ключевых терминов для обозначения тем может упростить интерпретацию выявленной темы и разрешить возможную неоднозначность. Так, для данной работы слово “документ”, встречающееся достаточно часто, чтобы подойти на роль термина, имеет широкую область употребления и способно относиться к различным предметным областям, например, документ инструкции в БД. Использование же ключевой фразы “документальный фильм” однозначно определяет одну из тем документа и сужает возможную предметную область статьи.

При этом стоит заметить, что в русском языке задача извлечения ключевых фраз является гораздо более сложной, чем, к примеру, в английском или немецком. Это связано с тем, что русский язык – флективный, то есть каждое слово в речи может быть представлено множеством различных словоформ. Обычные алгоритмы извлечения ключевых фраз, основанные на относительной частоте встречаемости n-грамм в документах, показывают низкий уровень точности извлечения. Каждую словоформу такие алгоритмы воспринимают как различные термины, и из-за этого частота встречаемости снижается в несколько раз. К примеру, если в тексте встретятся такие словосочетания, как “машинный код”, “машинного кода” и “о машинном коде”, классические алгоритмы каждой из этих биграмм присвоят разную частоту встречаемости, хотя на самом деле она должна быть общей и значительно выше, чем отдельные.

Существует несколько основных подходов к решению данной проблемы. Во-первых, для распознавания словоформ можно использовать словари, содержащие для каждого слова все его возможные формы [68]. В этом случае точность определения будет высокой для имеющихся в словаре слов (за исключением омонимии, например, слово “стекло” может являться словоформой глагола “стекать” либо существительным в именительном падеже). Однако очевидно, что применимость словарных алгоритмов ограничена предметной областью словаря. Был разработан модуль согласования словосочетаний на основе вышеперечисленных шаблонов, использующий для извлечения морфологической информации программу Mystem.

Другой подход к этой задаче – использование лексико-синтаксических шаблонов [69]. В [69] описана стратегия распознавания в заданном тексте фрагментов, соответствующих заданному лексико-синтаксическому шаблону, предложен язык записи шаблонов, позволяющий задавать лексические и грамматические свойства входящих в него элементов. К сожалению, основным недостатком методов, основанных на шаблонах, является их большая трудоемкость.

Проблему многословных терминов можно обойти, если использовать стемминг (нахождение основы слова) или лемматизацию (приведение слова к его начальной форме). Однако тогда возникает проблема с восстановлением изначальных словосочетаний. Так, в приведенном выше примере биграмма «тематическое моделирование» будет выглядеть после стемминга как «тематическ моделировани», а после лемматизации – как «тематический моделирование». Очевидно, такие биграммы не могут быть использованы в качестве ключевых фраз в научной статье или на веб-странице, и для дальнейшего использования нужно преобразовать их в изначальное словосочетание.

Для дальнейших рассуждений нам понадобятся несколько определений.

Будем называть словосочетание словосочетанием в начальной форме, если главное слово находится в словарной форме, а зависимые находятся в форме, обусловленной формой главного слова, а также видом связи в словосочетании.

Будем называть словосочетание лемматизированным, если каждое слово в нем находится в начальной форме при сохраненном порядке слов.

Будем называть согласованием словосочетания процесс преобразования его из лемматизированного вида в начальную форму.

В рамках данной работы были проведены эксперименты с двумя вариантами решения проблемы многословных терминов.

В качестве базового решения была выдвинута гипотеза, что в тексте статьи словосочетание обязательно встретится в своей начальной форме хотя бы один раз. Был разработан модуль поиска начальной формы в тексте для лемматизированного словосочетания. Эксперименты показали, что выдвинутая гипотеза была ошибочной – существуют

словосочетания с высокой частотой употребления в статье (что позволяет рассматривать их как кандидаты к использованию в качестве ключевых фраз), ни разу не встречающиеся в этой статье в начальной форме. Возможным вариантом улучшения такого подхода является использование большой базы статей для поиска начальной формы словосочетания, однако это значительно увеличит время работы программы.

Альтернативным решением проблемы согласования словосочетаний является использование для этого лексико-синтаксических шаблонов. Исследование многословных ключевых терминов, выбранных для статей авторами, позволило составить базовый набор, включающий в себя восемь шаблонов:

1. прилагательное в соответствующем роде, числе, падеже + существительное в начальной форме

Пример: *линейное уравнение*.

2. существительное_1 в начальной форме + существительное_2 в родительном падеже

Пример: *разработка системы*.

3. существительное_1 в начальной форме + прилагательное в соответствующем существительному_2 роде, числе, падеже + существительное_2 в родительном падеже

Пример: *гипотеза условной независимости*.

4. прилагательное_1 в соответствующем роде, числе, падеже + прилагательное_2 в соответствующем роде, числе, падеже + существительное в начальной форме

Пример: *вероятностная тематическая модель*.

5. существительное_1 в начальной форме + существительное_2 в родительном падеже + существительное_3 в родительном падеже

Пример: *определение тематики документа*.

6. прилагательное в соответствующем роде, числе, падеже + существительное_1 в начальной форме + существительное_2 в родительном падеже

Пример: *общая теория относительности*.

7. существительное_1 в начальной форме + существительное_2 в творительном падеже

Пример: *умножение столбиком*.

8. существительное_1 в начальной форме + существительное_2 в творительном падеже + существительное_3 в родительном падеже

Пример: *решение методом прогонки*.

Шаблоны (1) и (4), а также (2) и (5), могут быть обобщены до следующих шаблонов:

– прилагательное в соответствующем роде, числе, падеже * $n(n > 0)$ + существительное в начальной форме;

– существительное в начальной форме + существительное в родительном падеже * $n(n > 0)$.

Вопрос о полноте набора шаблонов терминов пока остается открытым. Однако предусмотрено возможное расширение набора шаблонов, и в случае увеличения их количества потребуются лишь минимальные изменения в модуле согласования словосочетаний.

Выделенные шаблоны удобно записать в терминах логики предикатов первого порядка. Рассмотрим словарь V – множество слов коллекции документов. Пусть $x, x_1, x_2, \dots, x_i, \dots, x_n$ – множество прилагательных из V ; $y, y_1, y_2, \dots, y_i, \dots, y_m$ – множество существительных из V . Для морфологических признаков введем следующие обозначения: $z1 = \{mal, fem, neu\}$ – содержит информацию о категории рода (мужской, женский, средний); $z2 = \{sin, plu\}$ – содержит информацию о категории числа (единственное, множественное); $z3 = \{nom, gen, dat, acc, ins, pre\}$ – содержит информацию о категории падежа (именительный, родительный, дательный, винительный, творительный, предложный). Далее введем четырехместные предикаты $A(x, z_1, z_2, z_3)$ для прилагательных и $N(x, z_1, z_2, z_3)$ для существительных. Теперь наши шаблоны многословных терминов можно записать в виде формул исчисления предикатов, т.е. в случае согласованных словосочетаний будут истинны формулы:

1. $MWE_1(x, p): A(x, z_1, z_2, nom) \wedge N(p, z_1, z_2, nom)$
2. $MWE_2(p_1, p_2): N(p_1, z_1^1, z_2^1, nom) \wedge N(p_2, z_1^2, z_2^2, gen)$
3. $MWE_3(p_1, x, p_2): N(p_1, z_1^1, z_2^1, nom) \wedge A(x, z_1^2, z_2^2, gen) \wedge N(p_2, z_1^2, z_2^2, gen)$
4. $MWE_4(x_1, x_2, p): A(x_1, z_1, z_2, nom) \wedge A(x_2, z_1, z_2, nom) \wedge N(p, z_1, z_2, nom)$
5. $MWE_5(p, y_2, y_3): N(p_1, z_1^1, z_2^1, nom) \wedge N(p_2, z_1^2, z_2^2, gen) \wedge N(p_3, z_1^3, z_2^3, gen)$
6. $MWE_6(x, y_1, y_2): A(x, z_1^1, z_2^1, nom) \wedge N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, gen)$
7. $MWE_7(y_1, y_2): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, ins)$
8. $MWE_8(y_1, y_2, y_3): N(y_1, z_1^1, z_2^1, nom) \wedge N(y_2, z_1^2, z_2^2, ins) \wedge N(y_3, z_1^3, z_2^3, gen)$

Обобщение шаблонов (1) и (4) теперь можно переписать в виде формулы:

$$\bigwedge_{i=1}^n A(x_i, z_1^i, z_2^i, nom) \wedge N(p, z_1, z_2, nom)$$

Обобщение шаблонов (2) и (5) запишем теперь в виде:

$$N(p_1, z_1^1, z_2^1, nom) \wedge \bigwedge_{j=2}^m N(p_j, z_1^j, z_2^j, gen)$$

2.4 Выводы по главе 2

В данной главе рассмотрены модели представления отдельно взятого текста и коллекции текстовых документов. Тематическое моделирование заключается в построении модели некоторой коллекции текстовых документов. В такой модели каждая тема представляется дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем.

Современные требования к тематическим моделям довольно разнообразны. Основное из них заключается в том, что тематические модели должны хорошо поддаваться интерпретации, конечному пользователю должно быть понятны причины выделения определенных тем в тексте и структура самих тем. Эта особенность является главным преимуществом тематических моделей перед набирающими популярность нейронными сетями. Кроме того, часто требуется, чтобы тематические модели учитывали разнородные данные, выявляли динамику тем во времени, автоматически разделяли темы на подтемы, использовали не только отдельные ключевые слова, но и многословные термины и т.д.

В данной главе также рассмотрена проблема согласования многословных терминов и предложены возможные пути ее решения.

Глава 3. Гибридный метод автоматического построения аннотаций научных текстов

Пусть текст T , очищенный после предварительной обработки, состоит из предложений

$$T = \bigcup_{k=1}^N S_k.$$

В нашем понимании задача автореферирования состоит в том, чтобы найти преобразование текста T в реферат \tilde{T} , такое, что

$$\Phi: T \rightarrow \tilde{T}, |\tilde{T}| < |T|.$$

Тогда алгоритм построения реферата, можно записать в виде последовательных этапов.

1. Предварительная обработка текста. На этапе предварительной обработки из исходного текста удаляются все изображения, таблицы, предложения с формулами, информация об авторах и библиографические ссылки. Авторские аннотации были убраны и отдельно сохранены, чтобы потом можно было оценить систему, путем сравнения результата с исходной аннотацией. Для лемматизации текста и построения морфологического словаря используется программа Mystem [70]. Программа лемматизирует слова, используя анализ контекста для снятия лексической неоднозначности, а также предоставляет морфологическую информацию (часть речи, род, число, падеж, склонение и др.) для каждого слова.

2. Построение тематических моделей, извлечение ключевых слов и многословных терминов. Первоначально строится униграммная модель текста, затем производится расширение модели многословными терминами. *Расширенной моделью* назовем тематическую модель, содержащую, помимо однословных терминов, термины, состоящие из нескольких слов (также называемые многословными терминами или ключевыми фразами). Такие модели лучше интерпретируемы для пользователя и точнее описывают предметную область документа, чем модели, состоящие только из униграмм (отдельных слов). В качестве алгоритма построения униграммных тематических моделей используется алгоритм ARTM в реализации библиотеки BigARTM. Для извлечения многословных терминов используется алгоритм RAKE, адаптированный для работы с текстами на русском языке.

3. Риторический анализ и построение квазиреферата. На этом шаге обнаруживаются предложения, содержащие дискурсивные маркеры, коннекторы и специальная лексика, характерная для научных текстов. К этим предложениям применяются определенные действия (более подробное описание см. в разделе 3.2). В результате получается квазиреферат: $\Phi(T, D, V) = T'$.

4. Оценка весов предложений. Для формирования аннотации подсчитываются веса предложений. Опишем эту процедуру подробнее.

Пусть S'_k – произвольное предложение квазиреферата

$$T' = \bigcup_{k=1}^{P_1} S'_k.$$

При вычислении веса каждого предложения квазиреферата учитывается наличие в этом предложении ключевых слов (или многословных терминов), дискурсивных маркеров и коннекторов, а также некоторых слов, которые характерны для научных текстов. Для извлечения из текстов многословных терминов используется алгоритм RAKE, разработанный для определения значимых n-грамм в английских текстах. В ходе создания системы мы адаптировали алгоритм RAKE для работы с текстами на русском языке.

В итоге вес каждого предложения вычисляется по следующей формуле:

$$SW = \frac{1}{L} \cdot \sum_{i=1}^L w_i + \frac{1}{M} \cdot \sum_{j=1}^M v_j + \frac{1}{N} \sum_{k=1}^N d_k, \text{ где}$$

$W = \{w_1, \dots, w_L\}$ – множество весов ключевых слов и многословных терминов, $|W| = L$;

$V = \{v_1, v_2, \dots, v_M\}$ – множество весов глаголов и существительных, характерных для научно-технических текстов, $|V| = M$;

$D = \{d_1, d_2, \dots, d_N\}$ – множество весов дискурсивных маркеров и коннекторов, $|D| = N$.

5. Выбор наиболее важных предложений. Из полученного набора предложений (см. п. 3) для аннотации отбираются только те предложения, вес которых (см. п. 4) превышает заданную пороговую величину β :

$$\tilde{T} = \bigcup_{k=1}^{N_1} \{S'_k : SW > \beta\},$$

где $\beta = 0,15$ является константой, которая определяется эмпирически; от нее зависит, насколько сильно будет сокращен текст.

6. Сглаживание – процедура преобразования текста, позволяющая получить связный текст из разрозненных фрагментов и при необходимости дополнительно сократить его. Например, в процессе сглаживания заменяются или удаляются некоторые слова или словосочетания и т.д.

3.1 Построение униграммных и расширенных тематических моделей

3.1.1 Выбор алгоритма тематического моделирования

Выбор методов тематического моделирования обусловлен наличием определенных особенностей. Для сравнения некоторые из них приведены в таблице 3.

Таблица 3 – Сравнение методов тематического моделирования

Название метода	Увеличение количества параметров модели с ростом числа документов	Применимость к большим наборам данных	Использование многословных терминов	Единственность и устойчивость решения
PLSA	да, есть линейная зависимость	нет	нет	нет
LDA	нет	да	нет	нет
ARTM	нет	да	нет	да
ARTM + RAKE	нет	да	да	да

Также для выбора базового алгоритма построения униграммных тематических моделей был проведен ряд экспериментов. Для проведения экспериментов была подготовлена коллекция текстов научных статей на русском языке на основе выложенных в открытом доступе архивов научных журналов¹. Статьи были очищены от формул, таблиц, рисунков и библиографических ссылок, аннотация и ключевые слова были удалены. Размер коллекции составляет около 260 текстов.

Для оценки результатов были выбраны следующие метрики, реализованные в библиотеке BigARTM и описанные в работе [67]: перплексия, разреженность матриц Φ и Θ , доля фоновых слов, мощность ядер тем, чистота ядер тем, контрастность ядер тем. Определения и пояснения содержатся в разделе 4.1 данной работы.

Первоначальные эксперименты показали, что LDA дает значительно худшие результаты перплексии по сравнению с PLSA и ARTM. В связи с этим дальнейшее сравнение проводилось только для двух последних алгоритмов при числе проходов по коллекции 100. Результаты представлены в таблице 4.

Таблица 4 – Сравнение алгоритмов PLSA и ARTM

Метрика	PLSA	ARTM
Перплексия	754.784	751.888
Разреженность матрицы Φ	0.769	0.769

¹ Программные продукты и системы. URL: <http://www.swsys.ru/>

Cloud of Science. URL: <https://cloudofscience.ru/>

Сибирский психологический журнал. URL: <http://journals.tsu.ru/psychology/>

Разреженность матрицы Θ	0.005	0.635
Доля фоновых слов	0.059	0.050
Средняя чистота ядер тем	0.370	0.364
Средняя контрастность ядер тем	0.787	0.788
Средняя мощность ядер тем	2085.000	2085.600

По результатам эксперимента, приведенным в таблице 4, можно увидеть, что ARTM показывает аналогичные либо лучшие результаты по сравнению с PLSA для всех метрик, за исключением средней чистоты ядер, где ухудшение незначительно. Более детальный анализ по сравнению методов тематического моделирования приведен в разделе 4.1. В совокупности с особенностями алгоритмов, представленными в таблице 3, было принято решение использовать алгоритм ARTM в реализации библиотеки BigARTM [63]. Благодаря своей универсальности и гибкости настройки параметров моделей ARTM позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. Этот метод гарантирует единственность и устойчивость решения. У ARTM не наблюдается увеличение количества параметров модели с ростом числа документов, поэтому он может применяться к большим наборам данных. Кроме того, предложенная нами модификация позволяет использовать не только однословные, но и многословные выражения, что, на наш взгляд, повышает интерпретируемость модели.

3.1.2 Извлечение многословных терминов

Для извлечения многословных терминов из текстов используется алгоритм извлечения ключевых слов RAKE (Rapid Automatic Keyword Extraction). Суть алгоритма описана в работе [71] и обобщенно состоит в следующем. На вход алгоритму подается текст, в нем алгоритм обнаруживает слова-кандидаты, которые представляют собой отдельно стоящие содержательные слова или последовательности таких слов. Для отдельно стоящих слов-кандидатов вычисляется вес по формуле:

$$Score(w) = \frac{Deg(w)}{Freq(w)}, \text{ где}$$

$Deg(w)$ – степень слова в графе слов текста;

$Freq(w)$ – частота слова в тексте.

Для последовательностей слов вес вычисляется как сумма весов входящих в него слов:

$$Score(w_1 \dots w_n) = \sum_{i=1}^n Score(w_i).$$

Данный алгоритм был разработан для применения в текстах на английском языке и показал довольно хорошие результаты. Поэтому в данной работе он был адаптирован для работы с русскими текстами вместе с алгоритмом ARTM.

Для определения списка ключевых слов для каждого документа изначально предполагалось использовать список наиболее часто встречающихся терминов (одно- и многословных) для каждой темы, к которой относится данный документ. Однако данный подход привел к тому, что из документа извлекались ключевые слова темы, а не самой статьи: для различных документов списки ключевых слов были очень похожи, а термины, которые должны быть ключевыми, исходя из текста статьи, не попадали в список из-за низкой частоты встречаемости. Для решения данной проблемы было предложено использовать TF-IDF – статистическую меру, оценивающую важность каждого слова для документа, в котором оно встречается [72].

Мера TF-IDF является произведением двух множителей:

1. TF (*term frequency*) – частота слова:

$$tf(t, d) = \frac{n_t}{\sum_k n_k},$$

где n_t – число вхождений слова t в документ d , $\sum_k n_k$ – общее число слов в документе.

2. IDF (*inverse document frequency*) – обратная частота документа:

$$idf(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|},$$

где $|D|$ – общее число документов в коллекции D , $|\{d_i \in D \mid t \in d_i\}|$ – число документов в коллекции D , в которых встречается слово t .

Тогда формула меры TF-IDF: $tf - idf(t, d, D) = tf(t, d) \times idf(t, D)$.

Наибольшее значение TF-IDF будут иметь слова, которые часто встречаются в данном документе, но редко встречаются в остальных документах коллекции.

Проблему многословных терминов можно обойти, если использовать стемминг (нахождение основы слова) или лемматизацию (приведение слова к его начальной форме). Однако тогда возникает проблема с восстановлением изначальных словосочетаний. Так, в приведенном выше примере биграмма «тематическое моделирование» будет выглядеть после стемминга как «тематическ моделировани», а после лемматизации – как «тематический моделирование». Очевидно, такие биграммы не могут быть использованы в качестве ключевых фраз в научной статье или на веб-странице, и для дальнейшего использования нужно преобразовать их в первоначальное словосочетание.

3.1.3 Алгоритм построения расширенных тематических моделей

Модуль построения расширенных тематических моделей написан на языке Python 3 с использованием библиотеки BigARTM. Используемые в системе алгоритмы из этой библиотеки были настроены таким образом, чтобы получить оптимальные результаты относительно различных метрик (перплексия, разреженность и др.) для научных текстов на русском языке.

Обобщенная схема работы модуля представлена на рисунке 6. Далее приведено подробное описание процесса построения расширенной тематической модели и извлечения ключевых фраз.

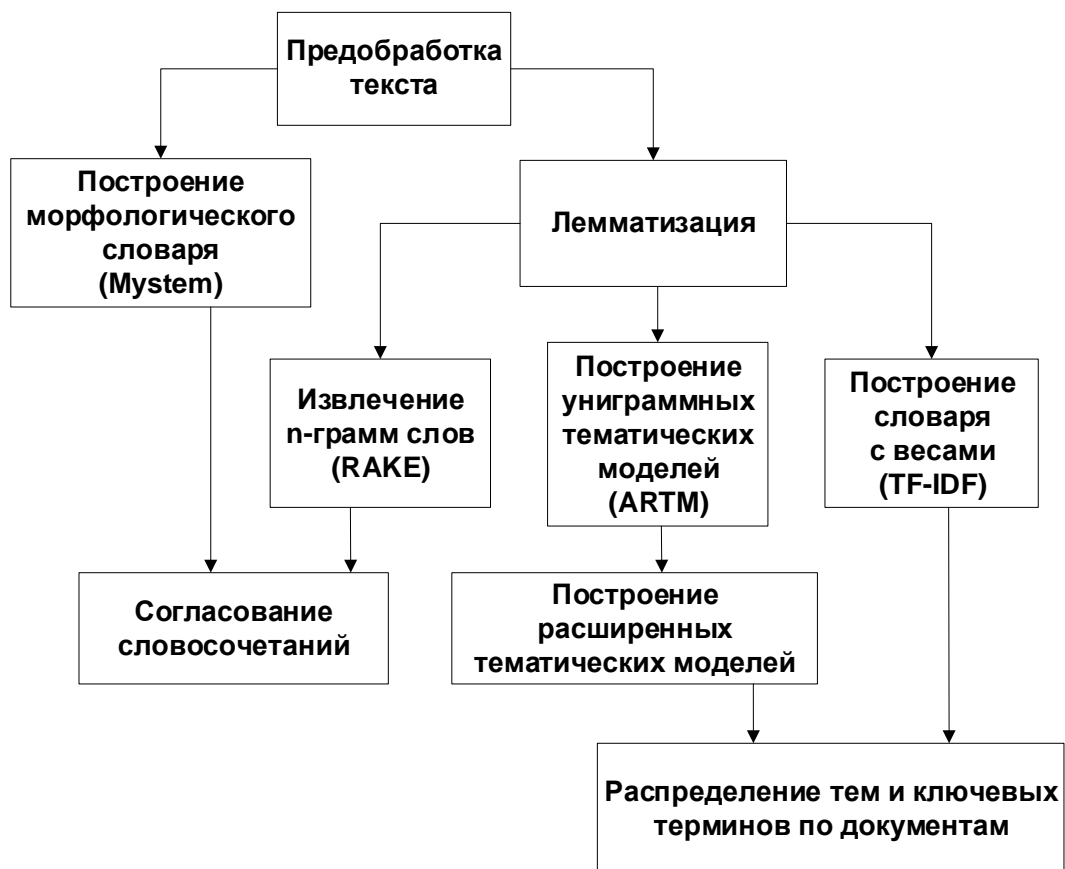


Рисунок 6 – Схема работы модуля построения тематических моделей

На вход модулю подаются лемматизированные словосочетания, которые сопоставляются с каждым шаблоном из набора. После определения требуемого шаблона словосочетание приводится в согласованный вид путем преобразования зависимых слов в форму, обусловленную формой главного слова и видом связи в словосочетании.

Опишем схему работы модуля как последовательность шагов.

Шаг 0. На вход модулю подается коллекция документов в формате .txt. Каждый документ должен быть представлен одним файлом, все документы помещены в одну директорию, путь к которой передается программе в качестве параметра.

Шаг 1. В модуле предобработки текста каждый документ очищается от специальных символов (отличных от кириллических и латинских букв), из документа удаляются стоп-слова (союзы, предлоги, частицы и др.), все слова приводятся к нижнему регистру. Далее строится корпус коллекции в формате последовательный Vowpal Wabbit: все документы коллекции помещаются в один файл, где каждая строка соответствует документу, первое слово в строке – название документа, остальные слова записаны в строке через одинарный пробел в порядке, соответствующем исходному тексту.

Шаг 2. Производится вызов программы Mystem, на вход которой подается файл с построенным на предыдущем этапе работы корпусом. Результатом работы является файл лемматизированного корпуса (формат, аналогичный полученному ранее корпусу, только каждое слово заменено его начальной формой), а также файл морфологического словаря, где каждой строке соответствует слово и описывающая его морфологическая информация.

Шаг 3. На лемматизированном корпусе производится поиск ключевых слов и n-грамм с помощью алгоритма RAKE.

Шаг 4. Найденные алгоритмом RAKE n-граммы преобразуются из лемматизированного вида в согласованный с использованием шаблонов и морфологического словаря, полученного на шаге 2.

Шаг 5. Для лемматизированного корпуса строится тематическая модель коллекции документов с использованием алгоритма ARTM. Параметры алгоритма можно как подобрать автоматически, так и использовать заранее вычисленные (так как подбор параметров – задача весьма трудоемкая и занимает значительное время).

Шаг 6. Полученная на шаге 5 тематическая модель расширяется с помощью многословных терминов, извлеченных из коллекции на шаге 3 и согласованных на шаге 4.

Шаг 7. Для каждого документа строится словарь TF-IDF: каждому слову в лемматизированном документе сопоставляется значение меры TF-IDF. Слова в словаре сортируются по убыванию значения меры.

Шаг 8. На основе матрицы распределения тем по документам каждому документу сопоставляется набор присутствующих в нем тем и их вероятностей (учитываются только темы, вероятность появления которых в данном документе превышает порог $\delta = \frac{1}{N_t}$, где N_t – количество тем в модели).

После этого сравниваются два множества: первые N_1 слов из отсортированного словаря TF-IDF и первые N_2 слов и словосочетаний для каждой темы, отсортированных по вероятности встретить этот термин в документе. Итоговыми ключевыми словами для темы документа будет пересечение этих множеств. N_1 и N_2 могут настраиваться; по умолчанию эти значения равны 100 и 300, соответственно. Такие значения параметров были подобраны эмпирическим путем, чтобы каждому документу в среднем соответствовало порядка 5-10 ключевых терминов.

Результатом работы программы является текстовый файл, содержащий следующую информацию:

- название исходного документа;
- список тем, для каждой из которых указана вероятность содержания ее в тексте как десятичная дробь от 0 до 1;
- список ключевых терминов для каждой темы.

Также для пользователя доступен файл с описанием тем, где каждой теме сопоставлено множество слов и словосочетаний с наибольшей вероятностью для этой темы.

Данный модуль показывает приемлемые результаты, а набор модулей покрывает значительную часть используемых в качестве ключевых фраз многословных терминов. Для улучшения результатов работы можно использовать как расширение набора используемых шаблонов, так и использование дополнительных способов согласования.

В дальнейшем планируется использовать модуль поиска начальной формы из базового подхода, модифицировав его для поиска всех вариантов заданного лемматизированного словосочетания, а затем применить морфологический анализатор для определения нужного числа существительных.

Схожим образом планируется устранить невозможность согласования словосочетаний, в которых присутствуют причастия. В лингвистике причастия считаются особой формой глагола, соответственно, Mystem при лемматизации приводит причастия к инфинитиву. Восстановить исходное причастие без использования дополнительной информации в этом случае не представляется возможным. Для согласования подобных словосочетаний необходимо извлекать априорную информацию о структуре слова из исходного текста, что планируется реализовать в дальнейшей работе над данной программой.

3.2 Риторический анализ и преобразования графов

Как уже говорилось ранее, в теории риторических структур можно определить два типа ЭДЕ. Один из них, называемый ядром, считается наиболее важной частью высказывания, другой, называемый сателлитом, поясняет ядро и считается вторичным. Ядро содержит

основную информацию, тогда как сателлит содержит дополнительную информацию о ядре. Сателлит часто непонятен без ядра. Между тем, выражения, в которых сателлит удален, могут быть поняты лишь в некоторой степени (см. рисунок 7).

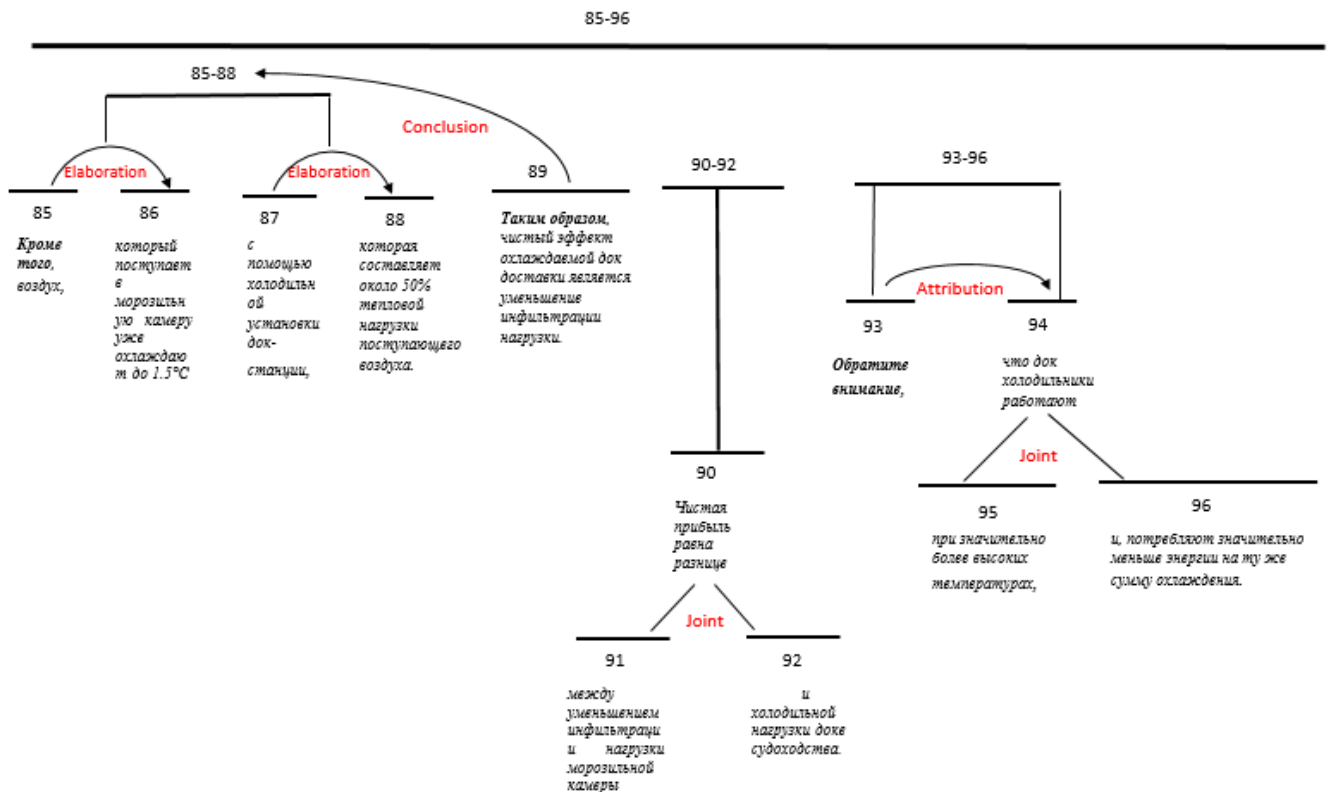


Рисунок 7 – Пример риторического анализа фрагмента текста

Маркеры (дискурсивные маркеры) – это слова или фразы, которые не имеют реального лексического значения, но вместо этого обладают важной функцией формировать разговорную структуру, передавая намерения говорящих при разговоре.

Коннекторы – группы слов, заменяющие маркеры и характеризующие определенные риторические отношения. Коннекторы обеспечивают связь между фразами, они показывают семантическую неполноту предложения. Например, «в связи с этим», «вместе с тем», «тем самым» и т. д.

3.2.1 Формальное описание преобразования текста

В предлагаемом подходе риторический анализ используется на этапе построения квазиреферата. Под квазирефератом понимается перечень наиболее значимых предложений текста. Упрощенно этот этап можно описать следующим образом. Сначала необходимо найти в тексте ядерные ЭДЕ. Далее следует преобразовать высказывания, содержащие эти ЭДЕ, так чтобы получился сокращенный текст, являющийся промежуточным между исходным текстом и

готовой аннотацией. В зависимости от разных маркеров и дискурсивных отношений эти преобразования будут разными. Для формального описания действий, выполняемых системой, было принято решение использовать логику предикатов первого и второго порядков. Рассмотрим пример.

Фрагмент текста: *Естественно описывать вычисление несколькими моделями. Кроме того, набор ограничений и утверждений о приложении может быть достаточно разнородным.*

Маркер: *Кроме того*

Название отношения: Elaboration

Для удобства введем следующие обозначения.

Предположим, что x – ядро; y – маркер; z – сателлит;

$S(x)$ – предикат для ЭДЕ, которая является ядром;

$S'(x)$ – предикат для ядра, которое начинается с заглавной буквы, т.е. находится в начале предложения;

$S(z)$ – предикат для ЭДЕ, которая является сателлитом;

$S'(z)$ – предикат для сателлита, который начинается с заглавной буквы;

y' – маркер с заглавной буквы;

$p()$ – символ пунктуации, аргументом может быть ".", ",", ":", ";".

Теперь приведенный пример может быть представлен в виде формулы исчисления предикатов: $S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.)$.

Согласно обозначениям, введенным в предыдущем разделе, для рассмотренного примера действия, выполняемые системой на этом этапе, могут быть записаны в таком виде:

$$S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.) \rightarrow S'(x) \wedge p(.) \wedge \neg(y' \wedge S(z) \wedge p(.)).$$

А именно, вначале надо найти маркер $y = \text{«кроме того»}$, потом необходимо удалить его вместе с сателлитом, оставив предыдущее предложение, которое является ядерным ЭДЕ.

Для предикатов, представленных выше, мы ввели специальные действия, которые выполняются для создания квазиреферата. Они зависят от некоторых глаголов, существительных, маркеров и коннекторов. Примеры маркеров, коннекторов и действий, связанных с ними, приведены в таблице 5.

Таблица 5 – Действия для маркеров и коннекторов

	Риторические отношения	Маркеры / коннекторы	Действия
1.	Elaboration	Кроме того	SAVE_DELETE
2.	Cause-Effect	Поэтому	DELETE_SAVE
3.	Contrast	Однако	SAVE_DELETE

4.	Elaboration	Например	SAVE_DELETE
5.	Evidence	Таким образом	DELETE_SAVE
6.	Restatement	Другими словами	SAVE_DELETE

Во время исследования была создана лингвистическая база знаний, состоящая из 140 маркеров и коннекторов, 120 существительных и 110 глаголов с весами, которые часто встречаются в научных и технических текстах. Всего было рассмотрено восемь действий. Ниже описаны некоторые действия.

MDELETE_SAVE: Это действие удаляет маркер предстоящего предложения и сохраняет предложение с заданным маркером (см. рисунок 8).



Рисунок 8 – Пояснение действия MDELETE_SAVE

MDELETE_SAVE: Конечный результат после операций (см. рисунок 9).

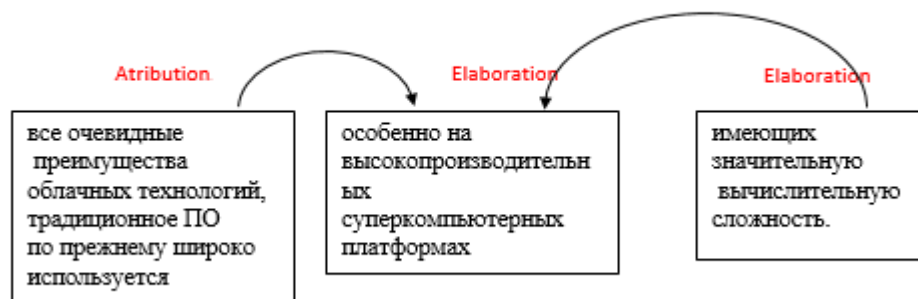


Рисунок 9 – Результат MDELETE_SAVE

SAVE_SAVE: Это действие полностью сохраняет предложение с заданным маркером и предыдущим предложением.

В сложноподчиненном предложении выделяется главное и придаточное предложение. В этом случае ЭДЕ более низкого уровня вложены в ЭДЕ более высокого уровня. Для описания действий с вложенными ЭДЕ удобнее использовать предикаты второго порядка. Чтобы проиллюстрировать, как текст преобразуется в случае вложенных ЭДЕ, приведем следующий пример.

*“**Кроме того**, воздух, который поступает в морозильную камеру уже охлаждают до 1.5°C с помощью холодильной установки док-станции, которая составляет около 50% тепловой нагрузки поступающего воздуха. **Таким образом**, чистый эффект охлаждаемой док доставки является уменьшение инфильтрации нагрузки. Чистая прибыль равна разнице между уменьшением инфильтрации нагрузки морозильной камеры и холодильной нагрузки доке судоходства. **Обратите внимание**, что док холодильники работают при значительно более высоких температурах (1.5°C вместо -23°C), и, потребляют значительно меньше энергии на ту же сумму охлаждения.”*

Для того чтобы записать пример в формальном виде, добавим следующие обозначения:

m – ядро в придаточном предложении;

n – сателлит в придаточном предложении;

$S(m)$ – предикат для ядра m ;

$S(m)$ – предикат для ядра m , начинающегося с заглавной буквы;

$S(n)$ – предикат для сателлита n ;

$S'(n)$ – предикат для сателлита n , начинающегося с заглавной буквы;

y – маркер.

Теперь преобразования с текстом можно описать следующим образом:

$$S'(z) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(z) \wedge p(.) \rightarrow \neg S'(El \wedge p(.)) \wedge S(m) \wedge p(.)) \wedge \neg S'(Ev \wedge p(.)) \wedge S'(S(m) \wedge p(.)) \wedge S'(x) \wedge p(.) \wedge \neg S'(Cont \wedge S(n) \wedge p(.)), \text{ где}$$

El = «*кроме того*»;

Ev = «*таким образом*»;

$Cont$ = «*обратите внимание*».

Следует отметить, что использование формализмов логики первого и второго порядка с данной целью пока недостаточно исследовано. В будущем, возможно, придется дополнить этот формализм, чтобы учитывался порядок следования элементов в тексте.

В данной работе риторический анализ используется для создания квазириферета из исходного текста. Алгоритм формирования квазириферата состоит из следующих шагов:

- 1) Искать маркеры Elaboration. Пронумеровать их как единое целое.
- 2) Удалять маркеры Background. Сохранять предыдущее предложение.
- 3) Для маркеров Contrast и Restatement соединить предложения и сократить их.
- 4) Предложения с маркерами Evidence, Concession, Cause-Effect, Purpose сохранять без изменений.

3.3 Операция сглаживания

Как уже отмечалось ранее, операция сглаживания – процедура преобразования текста, позволяющая получить связный текст из разрозненных фрагментов и при необходимости дополнительно сократить его. Например, в процессе сглаживания заменяются или удаляются некоторые слова или словосочетания и т.д. В таблице 6 приведены примеры предложений до сглаживания и после него.

Таблица 6 – Примеры сглаживания

До сглаживания	После сглаживания
<i>Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ, так как в некоторых ситуациях эпизоды могут быть настолько продолжительными, что задержки процесса обучения, связанные с необходимостью завершения эпизодов, будут слишком велики.</i>	<i>Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ.</i>
<i>Поскольку, как уже отмечалось, использование БСД позволяет анализировать лишь один из возможных диагнозов.</i>	<i>Выявлено что, использование БСД позволяет анализировать лишь один из возможных диагнозов.</i>

Для сглаживания предложений используются шаблоны двух типов: для удаления фрагментов предложений (в случае, когда аннотация получилась длиннее 250 слов) и для дополнения (в случаях, когда в аннотацию попал фрагмент незаконченного предложения). В тех случаях, когда требуется замена одного фрагмента предложения другим, сначала применяется подстановка в шаблон для удаления, затем подстановка в шаблон для дополнения. При этом важно, чтобы были выполнены определенные условия для выбора подходящих шаблонов.

Для дополнения использовались следующие типы шаблонов:

Введение;

Новизна (Применение | Актуальность | Эффективность | Особенность | Перспективность);

Цель;

Метод (Методика | Планирование | Методология | Модель | Стратегия | Подход | Оценка |
Определение | Формирование | Анализ | Проектирование);

Реализация;

Недостатки (Ошибки | Достоинства);

Заключение (Вывод | Итог | Результаты).

Тип шаблона «**Введение**» имеет вид $\langle X, Y_v, Y_n, Z \rangle$, где

X – добавляемый фрагмент. $X \in \{\text{«В статье»}, \text{«В работе»}, \dots\}$;

Y_v – глагол $V \in \{\text{«рассматриваются»}, \text{«рассматривается»}, \dots\}$;

Y_n – часть ядра, в состав которого входит существительное, характерное для научной лексики $N \in \{\text{«задачи»}, \text{«метод»}, \text{«способ»}, \text{«подходы»}, \dots\}$;

Z – оставшаяся часть предложения (сателлит).

Шаблоны типа «**Цель**» имеют несколько вариантов представления. Например, наиболее частый из них $\langle X_p, X_w, Y_v, KW, Z \rangle$, где

X_p – {Целью, Основной целью, Основным направлением, ...}

X_w – {данной работы, статьи, исследования, модели, ...}

Y_v – {является, играет, занимает, считается, ...}

KW – ключевые слова;

Z – оставшаяся часть предложения (сателлит с маркером или без маркера).

Тип шаблона «**Новизна**» имеет вид $\langle X, Y_n, Y_v, Z \rangle$, где

X – добавляемый фрагмент. $X \in \{\text{«Новизна»}, \text{«Новизна и перспективность»}, \dots\}$;

Y_n – существительное $N \in \{\text{«метода»}, \text{«алгоритма»}, \text{«подходов»}, \dots\}$;

Y_v – часть ядра, в состав которой входит глагол $V \in \{\text{«заключается»}, \text{«определяется»}, \dots\}$;

Z – оставшаяся часть предложения (сателлит).

В Приложении Б описаны рассмотренные нами шаблоны.

3.4 Разработка системы

В ходе выполнения данного исследования была разработана система построения аннотаций и тематических моделей, извлечения ключевых слов и фраз для текстов научных статей на русском языке. На рисунке 5 представлена блок-схема.

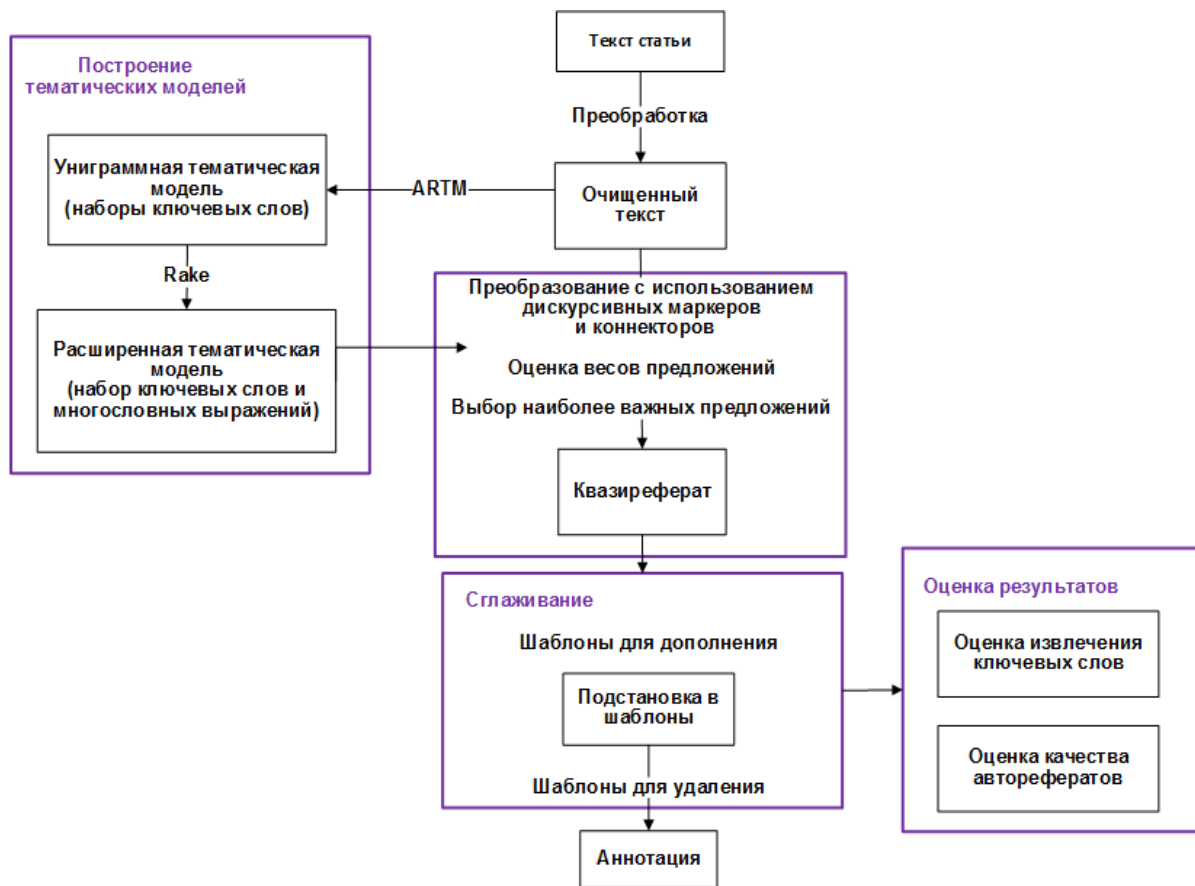


Рисунок 5 – Блок-схема системы Scientific Text Summarizer

Система реализована на языке Python3, также используется инструмент для работы с базами данных PostgreSQL. Используются внешние библиотеки Scikitlearn, Gensim, TensorFlow, NLTK, BigARTM, Flask и некоторые другие. В настоящее время программы функционируют в режиме веб-приложения. В Приложении А приведен фрагмент базы данных, содержащей маркеры и коннекторы, в Приложении В представлены промежуточные и итоговые результаты работы системы.

3.4 Применение предложенных методов для обработки текстов на тюркских языках

В данном разделе изложены методологические принципы применения предложенных методов для обработки текстов на тюркских языках, таких как казахский и турецкий. Эти языки относятся к типу агглютинативных языков. Агглютинативный язык – язык, имеющий строй, при котором доминирующим типом словоизменения является агглютинация («приклеивание») различных формантов (суффиксов или префиксов), причем каждый из них несет только одно значение. Такие языки обладают сложной и богатой морфологией.

Обычно слова в казахском и турецком языках состоят из основы и добавляемых к ней аффиксов (суффикс + окончание), которых бывает, по крайней мере, два или три. Вследствие особенностей словообразования уже на этапах морфологического и синтаксического анализа обнаруживаются семантические отношения между словами. В настоящее время над созданием систем морфологического и синтаксического анализа, ориентированных на агглютинативные языки, трудится много коллективов, в частности, в работах описаны различные подходы [73-76]. В большинстве работ авторы ограничиваются рассмотрением морфологического строения казахского или турецкого языков, проводят их сравнительный анализ. Исследования по синтаксису и семантике представлены в небольшом количестве.

3.4.1 Особенности морфологического анализа

Для казахского языка в данный момент нами было исследовано два подхода к автоматизации морфологического анализа. Один из них основан на правилах и реализован нами в виде самостоятельного приложения, в котором можно осуществлять стемматизацию существительных, прилагательных и глаголов [94]. В основу построения алгоритма морфологического анализа и синтеза, опирающегося на правила, положено разбиение всех слов на классы, определяющие характер изменения буквенного состава форм слов. Эти классы условно названы морфологическими. Изменения форм слов могут носить различный характер. Они могут быть связаны как с изменением формообразующих аффиксов слова, так и его основы (что в казахском языке бывает крайне редко: так, для существительных имеется 18 исключений, для глаголов – 352).

Морфологические классы слов делятся на два вида [77]: основоизменяемые классы, характеризующие систему изменения слов, и флективные классы слов. Флективные классы изменяемых слов выделялись на основе анализа их синтаксической функции и систем падежных, личных и родовых окончаний. Классы неизменяемых слов выделялись только по синтаксическому принципу. Общая морфологическая форма определения состава выглядит следующим образом:

түбір (корень) + жұрнақ (суффикс) + жалғау (окончание).

Принципиальным отличием морфологии казахского языка от морфологии, например, русского является наличие в казахском языке (как и в других тюркских языках) закона сингармонизма, в соответствии с которым аффиксы слова полностью определяются звуковым составом его основы. На основании анализа и грамматики казахского языка можно выделить следующие основные правила казахского языка.

Формально для существительных строится следующая модель образования словоформ. Обозначим через P_i следующие виды окончаний (аффиксов) для $i = 1, 2, 3, 4$:

P_1 – окончание множественного числа;

P_2 – притяжательное окончание;

P_3 – падежное окончание;

P_4 – личное окончание.

Возможны следующие комбинации окончаний существительных:

- 1) окончание множественного числа + притяжательное окончание ($P_1 P_2$);
- 2) окончание множественного числа + падежное окончание ($P_1 P_3$);
- 3) окончание множественного числа + личное окончание ($P_1 P_4$);
- 4) окончание множественного числа + притяжательное окончание + падежное окончание ($P_1 P_2 P_3$);
- 5) окончание множественного числа + притяжательное окончание + личное окончание ($P_1 P_2 P_4$);
- 6) притяжательное окончание + падежное окончание ($P_2 P_3$);
- 7) притяжательное окончание + личное окончание ($P_2 P_4$);
- 8) падежное окончание + личное окончание ($P_3 P_4$).

Для глаголов имеются следующие виды окончаний:

P_1 – окончание отрицания;

P_2 – окончание времени;

P_3 – личное окончание.

Возможны следующие комбинации окончаний глаголов:

- 1) окончание времени (P_2);
- 2) окончание времени + личное окончание ($P_2 P_3$);
- 3) окончание отрицания + окончание времени ($P_1 P_2$);
- 4) окончание отрицания + окончание времени + личное окончание ($P_1 P_2 P_3$).

Базовым алгоритмом при реализации подхода, основанного на правилах, является алгоритм Портера [78]. В зависимости от выполнения условий принимается решение, получена ли основа слова или требуется отсечение аффикса. Все необходимые для преобразований правила можно разделить на группы согласно флексивным классам.

В общих чертах алгоритм получения основ состоит из следующих этапов.

1. На вход поступает любая словоформа (глагол, существительное, прилагательное).
2. Начиная с последней буквы слова, происходит поиск по списку аффиксов.
3. Если данный аффикс найден, то он отсекается. Иначе, оставшаяся часть слова считается основой.

Основная проблема описываемого алгоритма – наличие в казахском языке слов, в которых последние буквы основы совпадают с одним из аффиксов. В этом случае алгоритм может отсечь больше, чем нужно. Единственный возможный механизм предотвращения таких ошибок – составление словаря основ, последние буквы которых совпадают с одним из аффиксов.

Были созданы словари, включающие в себя около 3500 аффиксов и их комбинаций (вариантов окончаний) для 14 флективных классов существительных и прилагательных, а также около 2000 глагольных аффиксов и их комбинаций для 17 флективных классов (некоторые сочетания аффиксов повторяются).

Второй подход к автоматизации морфологического анализа основан на грамматике связей и реализован в системе LGP (Link Grammar Parser). Результатом работы этой системы являются структуры, которые состоят из множества помеченных связей (коннекторов), соединяющих части слова попарно. Первоначально система LGP была создана как синтаксический анализатор. Подробное описание Link Grammar Parser можно найти в работе [79].

Основная идея грамматики связей позволяет наравне с синтаксической структурой предложения работать и с морфологией. При таком подходе можно рассматривать слова в качестве блоков с коннекторами. Существуют различные типы коннекторов; они могут указывать налево или направо. Левосторонний коннектор связывается с правосторонним коннектором того же типа. Вместе два коннектора образуют «связь». Правосторонний коннектор обозначается знаком «+», левосторонний – знаком «-».

На данный момент существуют подключаемые словари для английского, русского, персидского, арабского, немецкого, литовского, вьетнамского, индонезийского языков. Мы разрабатываем словарь для казахского и турецкого языков.

Связи для обозначения морфологических свойств слов

Связи, описывающие морфологические признаки слов, несут информацию как о словообразовании, так и о сочетании слов. Поскольку турецкий и казахский языки являются агглютинативными, образование новых слов и форм слов осуществляется последовательным присоединением аффиксов.

Выделяют [76, 80-82, 94] различные виды аффиксов для различных частей речи. Каждый вид отвечает за конкретный морфологический признак: число имени существительного, лицо глагола и т.д. Тогда каждой группе аффиксов сопоставим связь, при помощи которой он присоединяется к предыдущему аффиксу или основе. Теперь последовательное приписывание морфологических связей позволяет промоделировать процесс словообразования. Связи

являются направленными, а точнее, обратно направленными (от последнего аффикса к предыдущему, и далее к основе).

Например, глагол *читали* в турецком языке образуется следующим образом:

okuyorlar = *oku* + *yor* + *lar*, где

oku — основа;

yor — аффикс времени, означающий, что действие происходит в текущий момент;

lar — аффикс множественного числа.

Образование множественного числа существительных в турецком языке характеризуется наличием аффиксов *-lar* или *-ler*, присоединяющихся непосредственно к основе слова, т.е. эти аффиксы можно описать $\langle \text{lar, ler} \rangle$: {Nr-}. Аналогичные аффиксы есть в казахском языке: $\langle \text{лар, лер, дар, дер, тар, тер} \rangle$: {Nr-}. Тогда в описании основ слов в словаре должна присутствовать связь Nr+, как необходимая пара для Nr-.

Форма принадлежности существительных и местоимений турецкого языка характеризуется наличием аффиксов *-m, -im, -im, -um, -üm; -n, -in, -in, -un, -ün; -si, -si, -su, -sü, -i, -i, -u, -ü, -miz, -miz, -muz, -müz, -ımız, -ımız, -umuz, -ümüz, -niz, -niz, -nuz, -nüz, -ınız, -ınız, -unuz, -ünüz, -ları, -leri*. Похожая ситуация наблюдается в казахском языке. Множество существительных и местоимений с такими аффиксами в зависимости от лица описывается при помощи связей: Nr1-, Nr2-, Nr3-, Pr1-, Pr2-, Pr3-.

Аналогичным образом можно описать падежные аффиксы существительных: Nn — именительный; Ng — родительный; Nd — дательный; Na — винительный; Ni — творительный; Nl — местный (предложный); Nb — исходный. Некоторые из падежей почти полностью соответствуют русским падежам по значению, некоторые не имеют аналогий в русском языке и представляют собой особые формы.

Например, существительное *книгу* (*чью-то*) в турецком языке образуется следующим образом: *kitabını* = *kitab* + *ı* + *nı*, где *kitab* — основа, полученная от слова *kitab*; *ı* — аффикс принадлежности; *nı* — аффикс винительного падежа. Тогда согласно введенным обозначениям получаем следующий набор морфологических связей в словаре: $\langle \text{kitab} \rangle$: {Nr3+}; $\langle \text{ı} \rangle$: {Nr3-} & {Na+}; $\langle \text{nı} \rangle$: {Na-}.

В казахском языке, также как и в турецком, аффиксы обычно присоединяются в определенной последовательности. Сначала к основе присоединяется аффикс множественного числа, затем аффикс принадлежности, далее аффикс лица и в конце аффикс падежа. Руководствуясь этим правилом, будем дописывать связи в словаре.

Например, для существительного *друзьям* (*нашим*) в казахском языке: *достарымызға* = *дос* + *тар* + *ымыз* + *ға*, где *дос* — основа слова; *тар* — аффикс множественного числа; *ымыз* —

аффикс принадлежности; *ға* – аффикс дательного падежа; получаем представление в словаре: <дос>: {Nr+}; <тар>: {Nr-} & {Pr1+}; <ымыз>: {Pr1-} & {Nd+}; <ға>: {Nd-}.

Помимо этого, существуют словообразующие аффиксы, позволяющие получать прилагательные из существительных <лы, лі, ды, ді, ты, ті, сыз, сіз, дай, дей, тай, тей, лық, лік, дық, дік, тық, тік, ғы, гі, қы, кі>: {As-}, например, *ай (месяц) — айлық (ежемесячный)*. Или аффиксы, позволяющие образовывать глаголы из существительных и прилагательных <да, де, та, те, ла, ле, а, е, ар, ер, қар, кер, ғар, гер>: {Vna-}. Глагольные аффиксы, к тому же, требуют присоединения аффикса –у (который образует инфинитив глагола) или аффикса лица. Например глагол *бастау (начинать)* получается из существительного *бас (начало)* и имеет следующее описание в словаре: <бас>: {Vna+}; <та>: {Vna-} & {V+}; <у>: {V-}.

Существительные, образованные от глаголов, характеризуются наличием аффиксов <шы, ші, ғыш, гіш, қыш, кіш, ма, ме, ба, бе, па, пе>: {Sv-}, например, *оқу (учиться) — оқушы (ученик)*. Аффиксы, добавляемые к существительным, для образования новых существительных, <кер, гер, лас, лес, дас, дес, тас, тес, лық, лік, тық, тік, дық, дік, шы, ші>: {Ss-}, например, *ғарыш (космос) — ғарышкер (космонавт)*.

В казахском языке выделяют [82] следующие времена глагола. Каждое из времен характеризуется наличием определенных аффиксов. У субъективного прошедшего времени глагола имеются следующие аффиксы <ыпты, іпті>: {Vas+}; у результативного прошедшего времени <қан, ған, кен, ген>: {Var+}; у категорического прошедшего времени <ты, ті, ды, ді>: {Vas+}; у конкретного настоящего времени <п, ып, іп, а, е>: {Vr+}; у переходного будущего времени <ады, еді>: {Vft+}; у предположительного будущего времени <ар, ер>: {Vfs+}; у целенаправленного будущего времени <мақ, мек, пақ, пек>: {Vfg+}. На рисунке 10 приведен разбор предложения, содержащего глагол с аффиксом целенаправленного будущего времени: *Ол кешке болады. (Он будет вечером. He will be in the evening.)*

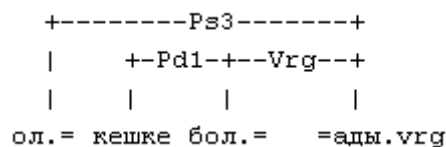


Рисунок 10 – Предложение, содержащее глагол с аффиксом целенаправленного будущего времени

3.4.2 Особенности синтаксического и риторического анализа

Синтаксические функции слов в предложении будем обозначать заглавными латинскими буквами. Для казахского и турецкого языков мы выделили следующие основные связи: AS – определение при подлежащем; АО – определение при дополнении; Е – обстоятельство при

сказуемом; J — соединяет послелог с существительным; OV — прямое дополнение при сказуемом; OJV — косвенное дополнение при сказуемом; S — соединяет подлежащее и сказуемое.

Если учитывать синтаксические функции слов в предложении, то каждой части речи можно сопоставить формулу из возможных связей. Рассмотрим пример структуры предложения на турецком языке. Имя существительное в предложении может выступать в роли подлежащего, к которому относятся определение и/или дополнение, сказуемое всегда будет справа: <N_S>: {AS-} & {OV+} & S+.

Кроме того, существительное может выполнять функцию дополнения, слева от которого также может быть определение, а справа может находиться послелог и сказуемое. Такая структура в общем случае будет описана формулой: <N_O>: {AO-} & {OV+} & {OJV+}.

Другой пример означает, что глагол может выступать в предложении в качестве сказуемого, слева от которого может быть подлежащее, дополнение (прямое или косвенное) или обстоятельство: <V_P>: {EI-} & {OV-} & {OJV-} & {S-}.

При этом обязательно в описании прилагательного должна присутствовать связь AI+, как необходимая пара для AI-, а в описании наречий — связь EI+, как необходимая пара для EI-. В противном случае связь не будет обнаружена.

На рисунке 11 видно, что при разборе предложения *Адамдар алма жеді.* (*Люди съели яблоко. People ate an apple.*) парсер определил 2 синтаксические (S3p, OV) и 2 морфологические (Np, Va3p) связи.

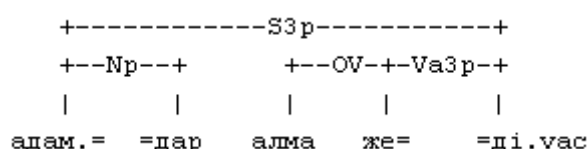


Рисунок 11 – Пример предложения с прямым дополнением

В следующем примере видно, что парсер определил 3 синтаксических (S3s, OJV, J) и 4 морфологических (Np, Va3s). Другой пример — предложение с косвенным дополнением — *Иттер мысықтардың артынан қуды.* (*Собаки гнались за кошками. Dogs chased the cats.*)

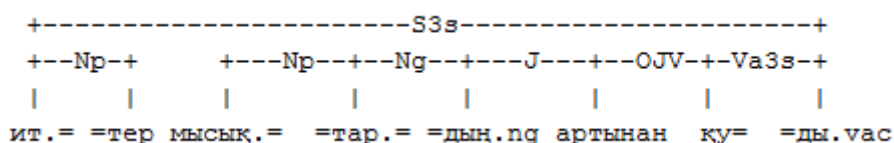


Рисунок 12 – Пример предложения с косвенным дополнением

Следует заметить, что синтаксические связи могут возникать не только между словами, но и между группами слов, например, в случае составного глагольного сказуемого, составного именного сказуемого, причастного оборота и т.д. [93]. К глаголу присоединяются определенные аффиксы (см. таблицу 7) и появляется дополнительный глагол.

Таблица 7 – Виды связей и аффиксы времен глаголов казахского языка

Вид времен глагола	Аффиксы	Связь
Субъективное прошедшее время	ыпты, іпті	{Vas+}
Результативное прошедшее время	кан, ған, кен, ген	{Var+}
Категорическое прошедшее время	ты, ті, ды, ді	{Vac+}
Конкретное настоящее время	п, ып, іп, а, е	{Vr+}
Переходное будущее время	ады, еді	{Vft+}
Предположительное будущее время	ар, ер	{Vfs+}
Целенаправленное будущее время	мақ, мек, пақ, пек	{Vfg+}

В турецком языке с окончаниями аффиксов глаголов в целом нами были выделены более 30 связей. Некоторые из них представлены в таблице 8. Так как аффиксы окончаний глаголов не повторяются и более точно распределены. Тем самым можно легко вывести их по связям, и с помощью четкого строения агглютинативных предложений, в том что глаголы и их окончания всегда стоят в конце и отделяемы, можно определить время, лицо и число, а также то, что они стоят в инфинитиве. В казахском языке же некоторые аффиксы повторяются и времена можно определить с помощью дополнительных слов времени.

Таблица 8 – Виды связей и аффиксы основных времен турецкого языка

Вид времен глагола	Аффиксы	Связь
Настоящее продолженное время	уор, іуор, іуор, ууор, үуор, öуор	{VPC+}
Субъективное прошедшее время	miş, miş, muş, müş	{VAS+}
Категорическое прошедшее время	dı, di, du, dü, ti, ti, tu, tü	{VAC+}
Категорическое будущее время	acak, ecek, aca, ece	{VFD+}
Настоящее будущее время	r, ır, ir, ur, ür, ar, er	{VFP+}
Прошедшее несовершенное время (past imperfect tense)	iyordu, iyordu, uyordu, üyordu, öyordu	{VAI+}

Все риторические отношения могут быть представлены парами трех видов:

–ядро-ядро $\langle N, N \rangle$ (Nucleus-Nucleus) – симметричное отношение;

–ядро-сателлит $\langle N, S \rangle$ (Nucleus-Satellite) – асимметричное отношение, в котором ядро предшествует сателлиту;

–сателлит-ядро $\langle S, N \rangle$ (Satellite-Nucleus) – асимметричное отношение, в котором сателлит предшествует ядру.

Как отмечено в работе [83] риторические отношения можно рассматривать как предикаты со свойствами, указывающими на определенные дифференцирующие признаки. Для некоторых риторических отношений удастся выделить маркеры. Исследование дискурсивных маркеров является одной из наиболее популярных областей дискурсивного анализа [84, 85]. К классу дискурсивных маркеров часто относят соединители-союзы (*когда, потому что, но* и т. д.). Помимо этого, сюда можно также отнести маркеры ментальных процессов говорящего (слова типа *вот, ну, так сказать*), маркеры контроля над ментальными процессами адресата (слова типа *понимаешь, видите ли*) и другие [86]. Заметим, что исследованиям дискурсивных маркеров в казахском языке в настоящее время не уделяется должного внимания, хотя принципы дискурсивного анализа не зависят от языка и могут одинаково успешно применяться как для русского, так и для других языков [123].

Ниже приведены маркеры, соответствующие им риторические отношения и примеры предложений с ними в таких же обозначениях, которые были введены ранее.

Пример 1.

Маркер: дегенмен (хотя)

Название отношения: Concession

Текст на казахском: Анизотропты және изотроптық жағдайларға арналған контурлық кернеулерді бөлу сапалық жағынан ұқсас, *дегенмен* кейбір сандық айырмашылықтар бар.

Текст на русском: Распределение контурных напряжений для анизотропного и изотропного случаев качественно подобно, *хотя* и имеется некоторое количественное различие.

Формульное представление примера на казахском языке: $S'(x) \wedge p(.) \wedge y \wedge S(z) \wedge p(.).$

Пример 2.

Маркер: сондықтан (поэтому)

Название отношения: Evidence

Текст на казахском: Жоғарыда келтірілген барлық параметрлердің әсерін қабылдауға және суффозия мен кольматацияны болжау мүмкін емес. *Сондықтан*, формулалар қатан

анықталған жағдайлар үшін пайдаланылады және құмды қондыру процесі ұңғыманың айналасында ғана сипатталады.

Текст на русском: Невозможно предсказать влияние всех вышеперечисленных параметров и предсказать кальцификацию и успокоение. *Поэтому* формулы используются для строго определенных ситуаций, и процесс миграции песка характеризуется только в области вокруг скважины.

Формульное представление примера на казахском языке: $S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.)$.

Как уже упоминалось ранее, случаи вложенных ЭДЕ, когда ЭДЕ более низкого уровня вкладываются в ЭДЕ более высокого уровня, удобнее описать при помощи предикатов второго порядка. Причем для каждого маркера вводится отдельный предикат. Краткий список подобных предикатов приведен ниже в таблице 9.

Таблица 9 – Предикаты для маркеров на казахском языке

Название отношения	Предикат	Маркер на казахском языке	Маркер на русском языке
Elaboration, Детализация	<i>El</i>	Сонымен қатар	Кроме того
Contrast, Контраст	<i>Cont1</i> <i>Cont2</i> <i>Cont3</i>	Алайда қарамастан Ескерту	Однако Несмотря на то, что Обратите внимание
Evidence, Обоснование	<i>Ev</i>	Осылайша	Таким образом
Cause-Effect, Причина	<i>CEf</i>	Сондықтан	Поэтому

Чтобы пояснить, как происходит преобразование текста в подобных случаях, приведем следующие примеры.

Пример 1.

Текст на казахском: *Сонымен қатар*, мұздатқыш камерасына енген әуе кіріс ауаның жылу жүктемесі шамамен 50% құрайды тоңазытқыш жүйесі док станциясын, пайдаланып °C 1,5 суыған. *Осылайша*, таза әсері $1 - 0,7 * 0,5 = 0,65$ ретінде шамамен 65% -ға қондыру жеткізу инфильтрация мұздатқыш жүктеме қысқарту суытылады. Таза табыс мұздатқыштың инфильтрация жүктемені азайту және тоңазытқыш жүктеме тасымалдау док арасындағы айырма болып табылады. *Ескерту* бұл док тоңазытқыштар әлдеқайда жоғары температура (1.5°C орнына -23°C) жұмыс, және салқындату сол сомаға әлдеқайда аз қуатты тұтынады.

Текст на русском: *Кроме того*, воздух, который поступает в морозильную камеру уже охлаждаются до 1.5°C с помощью холодильной установки док-станции, которая составляет около

50% тепловой нагрузки поступающего воздуха. **Таким образом**, чистый эффект охлаждаемой док доставки является уменьшение инфильтрации нагрузки морозилку примерно на 65% поскольку $1 - 0,7 * 0,5 = 0,65$. Чистая прибыль равна разнице между уменьшением инфильтрации нагрузки морозильной камеры и холодильной нагрузки доке судоходства. **Обратите внимание**, что док холодильники работают при значительно более высоких температурах (1.5°C вместо -23°C), и, потребляют значительно меньше энергии на ту же сумму охлаждения.

Формульное представление примера на казахском языке:

$$S'(z) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S(z) \rightarrow \neg S((El \wedge p(.)) \wedge S(z) \wedge p(.)) \wedge S'(\neg(Ev \wedge p(.)) \wedge S'(x) \wedge p(.)) \wedge S'(x) \wedge p(.) \wedge \neg S((Cont3 \wedge p(.)) \wedge S(z) \wedge p(.))$$

Пример 2.

Текст на казахском: *Алайда*, тіпті қалыпты заң үшін арифметикалық орта орташа мәнді бағалау болып табылмайды, ал медиана шығарындылардың болуында эмпирикалық орташа мәнді бағалауға мүмкіндік береді. **Сондықтан**, медиананы пайдалану регрессиялық тәуелділіктің параметрлерін баяулатуға мүмкіндік бергеніне **қарамастан**, орташа мәнді эмпирикалық бағалаулар медианы пайдалану регрессиялық тәуелділіктің параметрлерін баптау процедурасын жасайды.

Текст на русском: *Однако* даже для нормального закона среднее арифметическое не является робастной оценкой среднего значения, в то время как медиана позволяет оценивать эмпирическое среднее при наличии выбросов. **Поэтому** для построения параметрических регрессионных зависимостей также используются эмпирические оценки среднего при помощи медианы, **несмотря на то, что** использование медианы делает процедуру настройки параметров регрессионной зависимости более медленной.

Формульное представление примера на казахском языке:

$$S'(z) \wedge p(.) \wedge S(x) \wedge p(.) \rightarrow \neg(S(Cont1 \wedge S(z)) \wedge p(.)) \wedge S(\neg(CEf \wedge S(z) \wedge p(.)) \wedge Cont2) \wedge S'(x)) \wedge p(.).$$

Однако многие риторические отношения невозможно охарактеризовать наличием определенных дискурсивных маркеров. Кроме того, сами по себе маркеры являются не универсальными признаками, т. к. в разных естественных языках выражаются по-разному. Поэтому для более четкого описания риторических отношений необходимо выделить другие признаки. Так, в качестве признаков, могут быть взяты классифицирующие параметры, описанные М. Луверсом в работе [87]. В своей классификации М. Луверс выделил те параметры, которые наиболее часто используются для описания отношений когезии и когеренции. Когезия – структурная связность текста. Когерентность – содержательная связность текста. По большей части понятие «когерентность» применяется к содержательной

стороне текста, это организация содержания текста в целом, для которой особое значение имеет сама коммуникативная ситуация и набор знаний отправителя и получателя текста. Понятие «когезия» применяется к структурной организации текста, отвечает за присоединение единиц текста с помощью средств отдельных языковых уровней. Другими словами, когерентность – это свойство текста, а когезия – свойство элементов текста.

М. Луверс определил четыре вида параметров: тип отношения, полярность, направление и отражение связи в реальном мире.

Первый из параметров – тип отношений. Отношения могут быть трех типов: $TYPE = \{C, T, A\}$, C – каузальные (causal), T – темпоральные (temporal), A – аддитивные (additive). Каузальность содержит указание времени и причины, темпоральность – только времени, аддитивность не содержит каких-либо указаний.

Полярность отношений подразумевает их деление на положительные (positive) и отрицательные (negative): $POL = \{P, N\}$. Положительность означает, что ситуация, которая представлена первой, развивается во второй, присоединенной к ней ситуации. Негативность предполагает, что ожидаемая связь ситуаций прекращается, наоборот, присутствует противопоставление.

Направление может быть прямым (forward), обратным (backward) и двунаправленным (bi-directional) в зависимости от порядка упоминания событий в тексте: $DIR = \{B, F, BD\}$.

Отражение связи в реальном мире может быть рассмотрено на двух уровнях: между фактами и между речевыми актами. Первый уровень М. Луверс относит к семантике, второй – к прагматике. Семантические отношения разделяют на объектные (object-matter) и субъектные (subject-matter): $SEM = \{O, S\}$. Прагматические отношения разделяются на интенциональное (intentional), презентационное (presentational): $PRAG = \{IN, PR\}$.

Следует заметить, что последний параметр – отражение связей в реальном мире – является наименее исследованным в литературе и наиболее сложно поддается формализации. Поэтому было принято решение ограничиться рассмотрением только первых трех из перечисленных параметров. Последний планируется исследовать в дальнейшем.

Помимо указанных трех параметров некоторые риторические отношения удастся описать с помощью связей LGP, введенных в предыдущем разделе. Примеры описания представлены в таблице 10.

Таблица 10 – Описание отношений при помощи связей LGP и других признаков

Название отношения	Возможный маркер	Описание при помощи связей LGP	Тип	Полярность	Направление
--------------------	------------------	--------------------------------	-----	------------	-------------

Детализация, Elaboration	кроме того; более того (оған қоса)	$E+ \text{ or } (Xl- \& Xr+ \& (E+ \text{ or } E-)) \text{ or } (\{Xr+ \& \{Xl-\}\} \& OJV+) \text{ or } (\{Xr+ \& \{Xl-\}\} \& E-)$	add	pos	F
Причина, Cause-Effect	потому что (себебі)	$(J+ \text{ or } E+) \& (E- \text{ or } J- \text{ or } (\{Xr+ \& \{Xl-\}\} \& OJV+) \text{ or } (Xl- \& Xr+ \& E+))$	caus	pos	B / F
Условие, Condition	если..., то (егер ..., онда)	$(OV- \& \{Xc+ \& \{Xd-\}\} \text{ or } (OV- \text{ or } \{E+\}) \& ((\{Xl- \& Xr+\} \& E-) \text{ or } (\{Xr+ \& \{Xl-\}\} \& OJV+)))$	caus	pos	F
Уступка, Concession	несмотря на (оған қарамастан)	$(J+ \text{ or } E+) \& (E- \text{ or } (Xl- \& Xr+ \& E+) \text{ or } (\{Xr+ \& \{Xl-\}\} \& OJV+))$	add	neg	F
Уступка, Concession	хотя (дегенмен)	$(OV- \& ((\{Xr+ \& \{Xl-\}\} \& OJV+) \text{ or } (\{Xl- \& Xr+\} \& E-)))$	caus	neg	B / F
Цель, Purpose	чтобы (үшін)	$((J+ \text{ or } E+) \& (OV- \text{ or } MVI-)) \text{ or } (J+ \& (Xr- \text{ or } Xl+))$	caus	pos	F

Сформулируем следующие утверждения, связывающие выделенные признаки.

Утверждение 1. Аддитивность невозможна для обратного направления.

Пусть $TYPE = \{C, T, A\}$ и $DIR = \{B, F, BD\}$.

Тогда, если $R \in A$, то $R \in F$ или $R \in BD$, но $R \notin B$.

Утверждение 2. Не существует двунаправленной каузальности.

Пусть $TYPE = \{C, T, A\}$ и $DIR = \{B, F, BD\}$.

Тогда, если $R \in C$, то $R \in B$ или $R \in F$, но $R \notin BD$.

Утверждение 3. Не существует отрицательной двунаправленной темпоральности.

Пусть $TYPE = \{C, T, A\}$, $DIR = \{B, F, BD\}$ и $POL = \{P, N\}$.

Тогда, если $R \in T$ и $R \in BD$, то $R \in P$ и $R \notin N$.

Доказательство этих утверждений непосредственно следует из определений множеств $TYPE$, DIR и POL .

На основе предложенных методов был создан инструмент для определения риторических отношений [92]. С помощью него было решено провести эксперимент по обнаружению отношений, характеризующих тексты научно-технической тематики. В ходе эксперимента было проанализировано 168 статей на русском языке, средняя длина которых 7-12 страниц. На казахском языке была собрана коллекция из 207 статей. В эксперименте

рассматривалось в общей сложности 9 отношений и около 40 маркеров на каждом языке. Распределение рассмотренных риторических отношений в текстах показаны в таблице 11.

Таблица 11 – Распределение отношений для казахского и русского языков

	Название отношения	Кол-во примеров на казахском языке	Кол-во примеров на русском языке
1.	Детализация, Elaboration	631	629
2.	Контраст, Contrast	291	1112
3.	Обоснование, Evidence	223	802
4.	Переформулировка, Restatement	761	1274
5.	Причина, Cause-Effect	759	2123
6.	Сравнение, Comparison	92	77
7.	Уступка, Concession	1492	1329
8.	Фон, Background	1854	3458
9.	Цель, Purpose	3584	671

В результате можно сделать вывод, что для обоих языков научно-технические тексты в большей мере характеризуются следующими отношениями: *Переформулировка (Restatement)*, *Причина (Cause-Effect)*, *Уступка (Concession)*, *Фон (Background)*, *Цель (Purpose)*. Значит, в дальнейших исследованиях научно-технических текстов имеет смысл детальнее изучить именно их. Для обоих языков в сравнительно небольшом количестве представлены *Детализация (Elaboration)* и *Сравнение (Comparison)*.

3.5 Выводы по главе 3

В данной главе описана система автореферирования и используемые в ней методы, модели и алгоритмы. Обобщенный алгоритм автоматического построения реферата состоит из следующих этапов:

- 1) Предварительная обработка текста;
- 2) Построение тематических моделей (униграммной и расширенной);
- 3) Риторический анализ и формирование квазиреферата;
- 4) Оценка весов предложений;
- 5) Выбор наиболее важных предложений;
- 6) Сглаживание полученного текста.

Выбор метода тематического моделирования осуществлялся на основе дополнительного эксперимента по сравнению методов. В результате было принято решение использовать алгоритм ARTM в реализации библиотеки BigARTM [63]. Благодаря своей универсальности и гибкости настройки параметров моделей ARTM позволяет комбинировать тематические модели. Этот метод гарантирует единственность и устойчивость решения. У ARTM не наблюдается увеличение количества параметров модели с ростом числа документов, поэтому он может применяться к большим наборам данных.

Кроме того, предложенная нами модификация позволяет строить расширенные тематические модели, которые включают в себя не только однословные, но и многословные ключевые термины. Такие модели, на наш взгляд, легче интерпретируются пользователем и точнее описывают предметную область документа, чем модели, состоящие только из униграмм (отдельных слов). Для извлечения многословных терминов используется алгоритм RAKE, адаптированный для работы с текстами на русском языке.

В данной главе также предпринята попытка формально описать свойства риторических отношений и выполняемые преобразования с текстом на основе риторического анализа. Кроме того, описана процедура сглаживания, позволяющая получить связный текст из разрозненных фрагментов.

Помимо этого, в данной главе изложены методологические принципы применения предложенных методов для обработки текстов на тюркских языках, таких как казахский и турецкий. Описаны особенности автоматизации морфологического и синтаксического анализа языков такого строя.

Описанная в данной главе система реализована на языке Python3, также используется инструмент для работы с базами данных PostgreSQL. Используются внешние библиотеки Scikitlearn, Gensim, TensorFlow, NLTK, BigARTM, Flask и некоторые другие. В настоящее время программы функционируют в режиме веб-приложения. Исходные тексты разработанных программ пока не доступны в сети Интернет. В Приложении А приведен фрагмент базы данных, содержащей маркеры и коннекторы, в Приложении В представлены промежуточные и итоговые результаты работы системы.

Глава 4. Оценка эффективности разработанных методов

Выбор базового алгоритма построения униграммных тематических моделей осуществлялся на основе ряда экспериментов. Как уже упоминалось ранее, для проведения экспериментов была подготовлена коллекция текстов научных статей на русском языке на основе выложенных в открытом доступе архивов научных журналов². Статьи были очищены от формул, таблиц, рисунков и библиографических ссылок; аннотация и ключевые слова были удалены. Размер коллекции составил около 260 текстов.

4.1 Оценка тематических моделей и качества извлечения ключевых терминов

Для оценки и выбора тематических моделей были выбраны следующие метрики, реализованные в библиотеке BigARTM и описанные в работе [67]:

1. Перплексия
2. Разреженность матриц Φ и Θ
3. Доля фоновых слов
4. Мощность ядер тем
5. Чистота ядер тем
6. Контрастность ядер тем

Перплексия является наиболее распространенным критерием, используемым для оценивания тематических моделей. Это мера несоответствия модели $p(w/d)$ терминам w , наблюдаемым в документах d коллекции D , определяемая через логарифм правдоподобия:

$$P(D; p) = \exp\left(-\frac{1}{N}L(D, \Phi, \Theta)\right) = \exp\left(-\frac{1}{N}\sum_{d \in D}\sum_{w \in d} n_{dw} \ln p(w|d)\right)$$

Численное значение перплексии не имеет интерпретации и позволяет лишь сравнивать алгоритмы между собой. Чем меньше эта величина, тем лучше модель p предсказывает появление терминов w в документах d коллекции D .

Разреженность матриц Φ и Θ – доля нулевых элементов в матрицах распределения документов по темам и слов по темам. Для хорошей тематической модели эта величина будет

² Программные продукты и системы. URL: <http://www.swsys.ru/>

Cloud of Science. URL: <https://cloudofscience.ru/>

Сибирский психологический журнал. URL: <http://journals.tsu.ru/psychology/>

высокой, если же она приближается к нулю, значит, не было выделено ни одной хорошей предметной темы: в каждом документе присутствует большинство тем модели и/или в каждой теме присутствует большинство слов коллекции.

Доля фоновых слов выражается формулой:

$$\frac{1}{N} \sum_{d,w} n_{dwt},$$

где N – количество слов в коллекции, n_{dwt} – число вхождений слова w в каждый документ d , относящийся к теме t . Значение этой метрики может принимать значения от 0 до 1, причем значения, близкие к 1, свидетельствуют о вырождения тематической модели, к примеру, в результате чрезмерного разреживания.

Лексическим ядром темы называется множество слов, отличающих данную тему от остальных, т.е. множество $W_t = \{w \in W \mid p(t|w) > \delta\}$.

На основе ядра темы строятся следующие две оценки.

Чистота ядра темы – суммарная вероятность слов ядра:

$$\sum_{w \in W_t} p(t|w)$$

Эта мера показывает, насколько хорошо тема описывается своим ядром. Чем выше значение чистоты, тем лучше.

Контрастность ядра темы – средняя вероятность встретить слова ядра в конкретной теме:

$$\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$$

При большой контрастности тема однозначно угадывается по своему ядру, при малой контрастности тема размывается, становится нечеткой.

На графике (см. рисунок 13) представлена зависимость перплексии от числа итераций (проходов по коллекции):

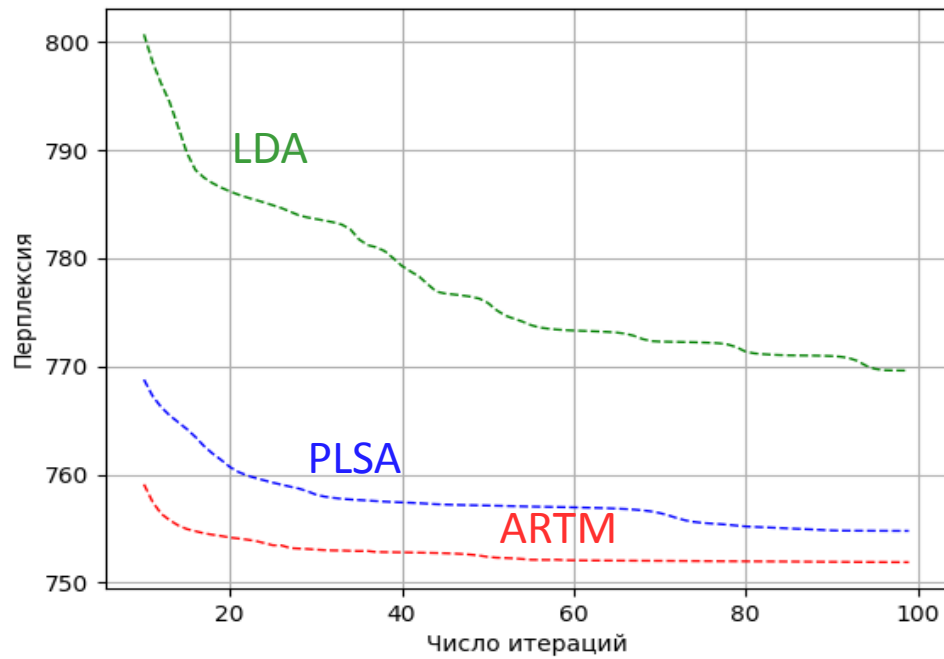


Рисунок 13 – Сравнение перплексии LDA, PLSA и ARTM

По графику видно, что LDA показывает значительно худшие результаты по сравнению с PLSA и ARTM. В связи с этим дальнейшее сравнение проводилось только для двух последних алгоритмов при числе проходов по коллекции 100. Результаты представлены на графиках (см. рисунки 14-19) и в таблице 4.

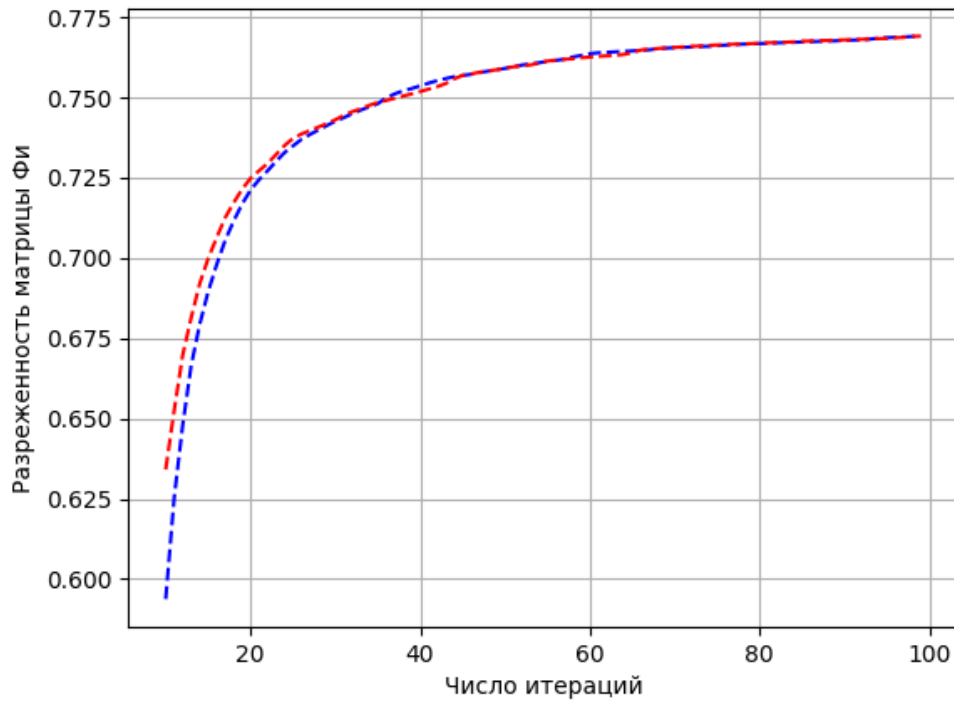


Рисунок 14 – Сравнение разреженности матрицы Ф

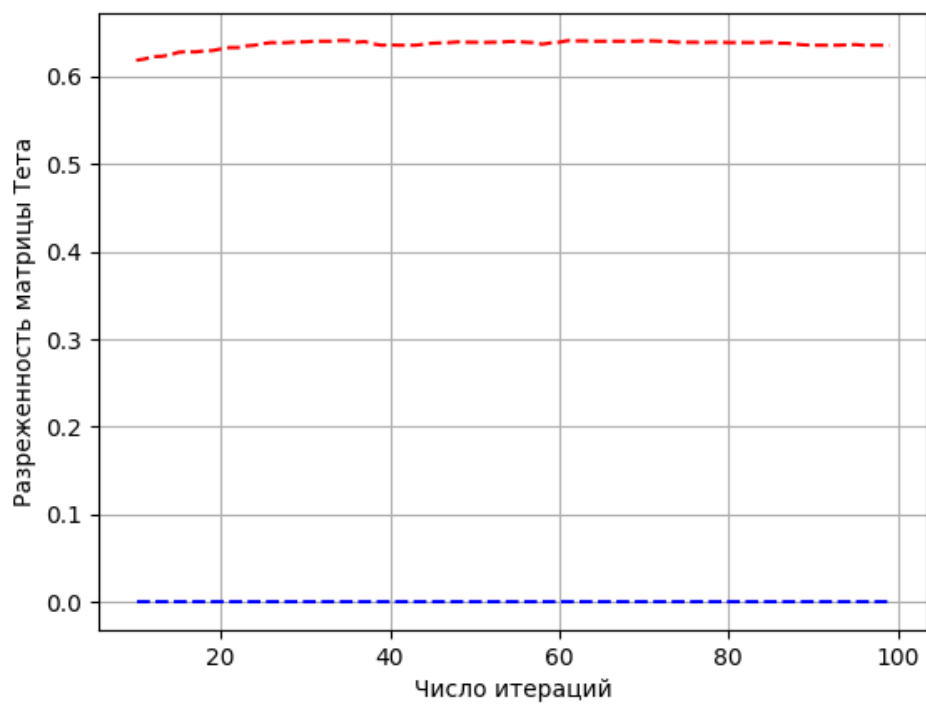
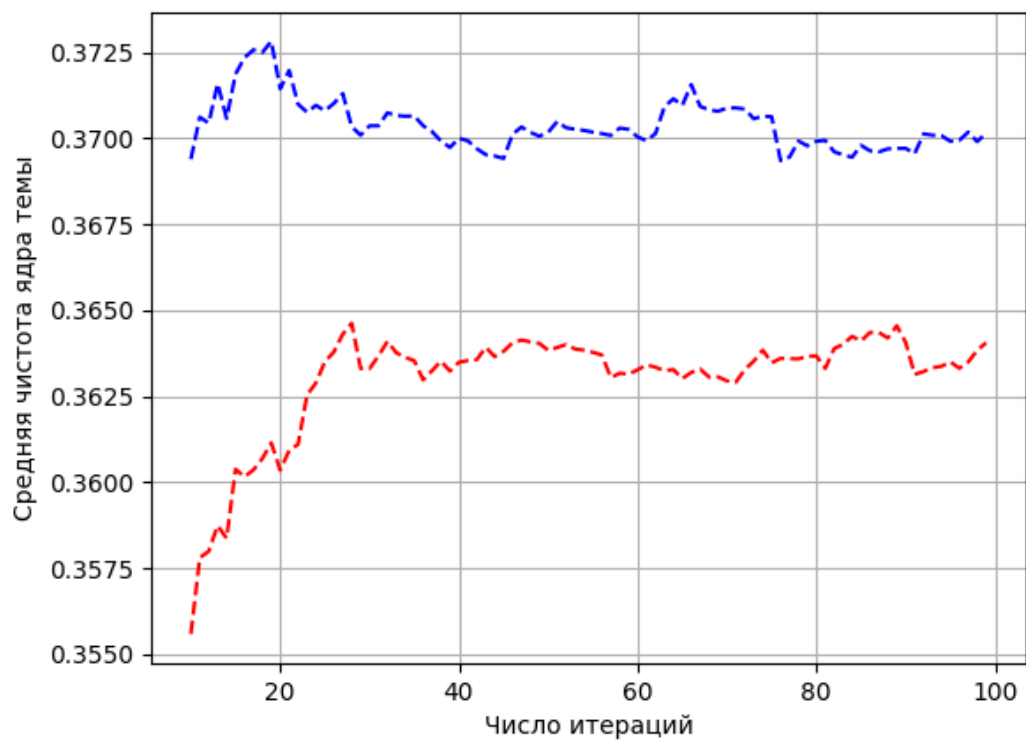
Рисунок 15 – Сравнение разреженности матрицы Θ 

Рисунок 16 – Сравнение средней чистоты ядер тем

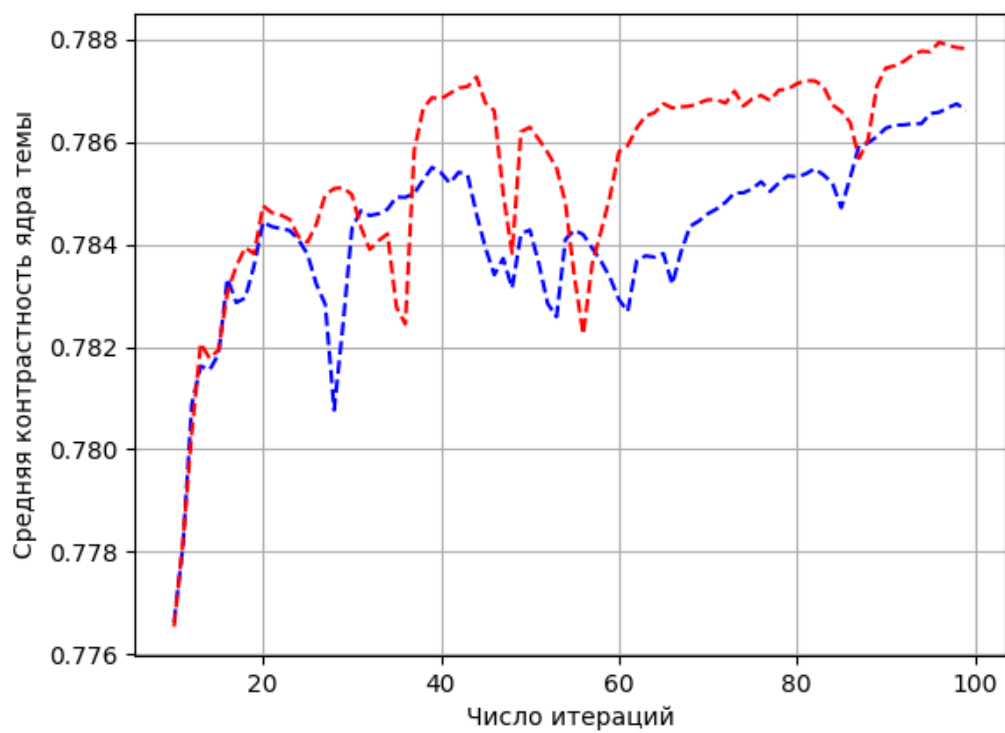


Рисунок 17 – Сравнение средней контрастности ядер тем

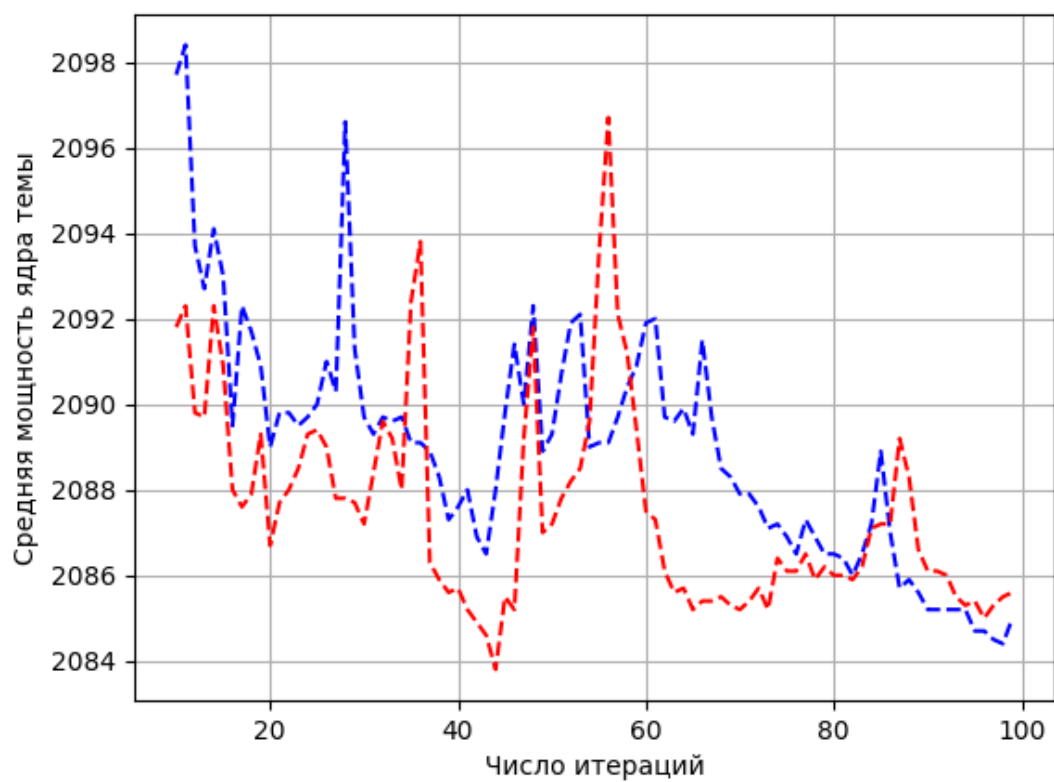


Рисунок 18 – Сравнение средней мощности ядер тем

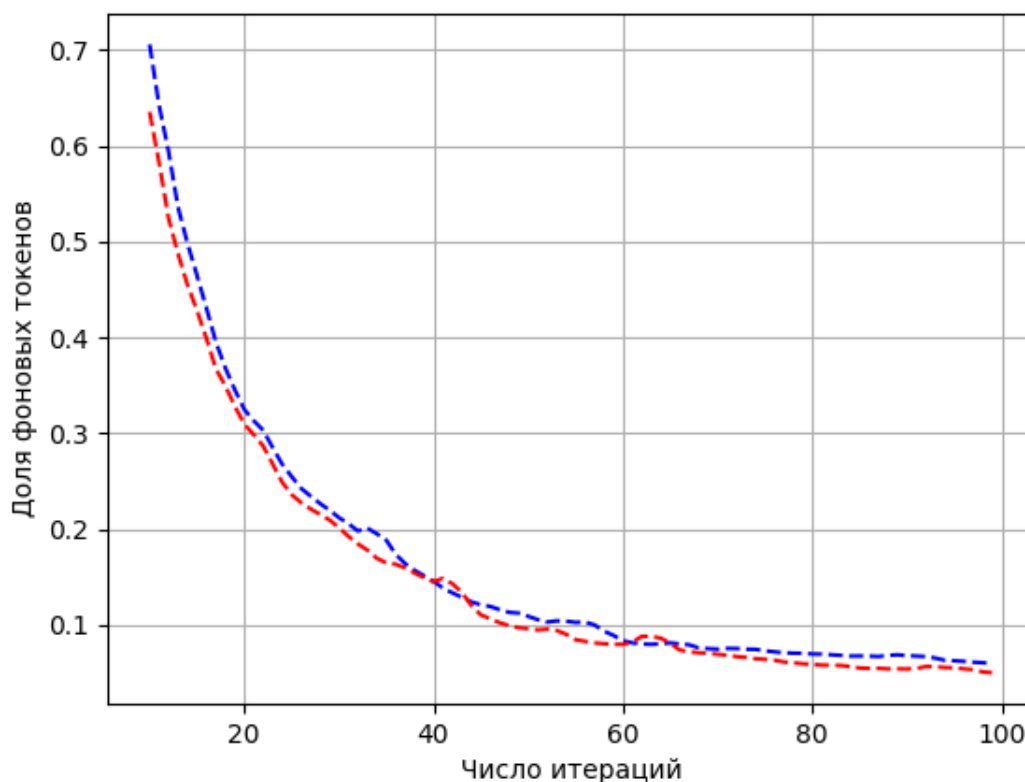


Рисунок 19 – Сравнение доли фоновых слов

Далее приведены несколько примеров работы алгоритма для различных документов разных тематик. Некоторые из наиболее частотных слов и фраз для первых пяти тем расширенной тематической модели коллекции представлены в таблице 12.

Таблица 12 – Расширенная тематическая модель коллекции научных статей

№ темы	Описание темы
Тема 1	'алгоритм', 'решение', 'задача', 'значение', 'вершина', 'значение параметра', 'время распознавания', 'класс объекта', 'обработка информации', 'алгоритм поиска', 'вершина графа', 'изображение объекта', 'граница решения', 'задача поиска', 'граф решения'
Тема 2	'метод', 'данные', 'алгоритм', 'классификация', 'текст', 'слово', 'классификатор', 'обучение', 'значение параметра', 'класс объекта', 'множество признака', 'представление текста', 'процесс обучения', 'метод классификации', 'построение модели', 'задача классификации', 'качество классификации', 'обучение классификатора', 'классификация текста'
Тема 3	'человек', 'ребенок', 'психологический', 'группа', 'отношение', 'стратегия'

	восприятия', 'процесс формирования', 'образ мира', 'группа испытуемых', 'уровень развития', 'респондент группы', 'развитие ребенка'
Тема 4	'система', 'управление', 'процесс', 'модель', 'требование', 'разработка', 'система управления', 'орган управления', 'процесс разработки', 'модель прогнозирования', 'критерий эффективности проекта', 'этап прогнозирования', 'критерий эффективности', 'эффективность проекта'
Тема 5	'исследование', 'отношение', 'испытуемый', 'элемент', 'диагностический', 'результат исследования', 'значение параметра', 'удовлетворенность отношения', 'процесс формирования', 'поиск решения', 'вид деятельности', 'группа испытуемых', 'удовлетворенность брака', 'формирование религиозности'

По представленным в таблице результатам можно отметить, что темы из разных предметных областей (технические науки и психология) очень хорошо различимы в тематической модели. При этом граница между более узкими темами не настолько четкая: если тема 4 довольно хорошо интерпретируема как отдельная предметная область, связанная с управлением проектами и процессом разработки, темы 1 и 2 обе связаны с классификацией и распознаванием, а темы 3 и 5 связаны с психологической диагностикой. При этом важно заметить, что в теме 5 многословные термины («удовлетворенность отношения», «формирование религиозности» и т.д.) улучшают интерпретируемость темы как относящуюся к психологии, тогда как термины «исследование», «испытуемый» являются более общими.

В таблице 13 представлены извлеченные программой ключевые термины для нескольких научных публикаций.

Таблица 13 – Ключевые слова и фразы

Название документа	Выделенные ключевые термины
1. Алгоритм детектирования объектов на фотоснимках с низким качеством изображения	объект, класс, изображение, набор, автокодировщик, обучение, объект, класс, набор, изображение, слой, пиксел
2. Проектирование интерфейса программного обеспечения с использованием элементов искусственного	программный, пользователь, система управления, уровень развития, нечеткий, интерфейс, характеристика, эксперт,

интеллекта	система управления
3. Родительское отношение как фактор формирования религиозности личности	ребенок, отношение, родитель, формирование, религиозность, религиозный, религия, семья, родительский, решение задачи
4. Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей	прогнозирование, состояние, клиент, классификатор, заемщик, решение задачи, дерево решения
5. Разработка системы хранения ансамблей нейросетевых моделей	данные, модель, набор, ансамбль, ряд, преобразование, хранение, нейросетевой, оценка качества, процесс формирования, классификация текста

Результат извлечения ключевых терминов оценивался при помощи стандартных метрик: точности, полноты и F-меры.

Точность – показатель того, сколько положительных решений правильные:

$$Precision = \frac{TP}{TP + FP}$$

Полнота – показатель того, сколько всего ключевых слов найдено:

$$Recall = \frac{TP}{TP + FN}$$

F-мера – представляет собой гармоническое среднее между точностью и полнотой:

$$F_{measure} = \frac{2 \cdot P \cdot R}{P + R}$$

В таблице 14 содержится сравнение результата работы нашей системы и других систем, находящихся в открытом доступе.

Таблица 14 – Оценка качества извлечения ключевых слов

Система	Метод	Точность, %	Полнота, %	F-мера, %
Open Text Summarizer (2016)	статистический	3	25	5
t-conspectus (2016)	TF-IDF	4	7	5
Scientific Text Summarizer (2018)	предложен в данной работе	44	37	40

Можно утверждать, что извлеченные ключевые слова и фразы соответствуют содержанию статей и хорошо определяют предметную область исследований. При этом можно заметить, что в некоторых случаях они дают большее представление о содержании публикации, чем ее название: например, ключевая фраза «дерево решения» дает понять, что в качестве алгоритма машинного обучения в четвертой статье использовались деревья решений, а ключевая фраза «классификация текста» в статье 5 указывает, что ансамбли нейросетевых моделей здесь использовались для классификации текста (а не, например, только изображений).

4.2 Оценка результатов реферирования

Оценка качества автореферирования осуществлялась на собранной нами коллекции по находящимся в открытом доступе материалам международного научно-практического журнала «Программные продукты и системы» за 2010-2018 гг. Объем этой коллекции составил 1200 статей.

Мы оценивали результаты несколькими способами: при помощи ROUGE, RAV, экспертной оценки, точности, полноты и F-меры, т.к. согласно [88] пока не существует общепринятого эффективного способа автоматической оценки систем автореферирования. Далее в этом разделе подробно приведены полученные нами результаты.

4.2.1 Метрика Rouge

Метрика ROUGE основана на подсчете количества соответствующих текстовых элементов, например, n-грамм или предложений [89]. В метрике ROUGE в случае подсчета совпадающих предложений текст аннотации рассматривается как последовательность предложений. Основная идея состоит в том, что чем длиннее самая длинная общая подпоследовательность LCS двух предложений в сравниваемых аннотациях, тем более похожими считаются эти две аннотации. Как правило, используют F-меру на основе LCS для оценки сходства между двумя величинами X длиной m и Y длиной n , считая, что X является образцом для сравнения, а Y – просматриваемый элемент. Точность, полнота и F-мера согласно ROUGE определяются следующим образом:

$$P_{lcs} = \frac{LCS(X,Y)}{n}, R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Здесь $LCS(X, Y)$ – длина самой длинной общей подпоследовательности X и Y , а $\beta = P_{lcs}/R_{lcs}$.

Были получены следующие значения метрики ROUGE: точность 32,8 %, полнота 59,04 %, F-мера 34,47 %. К сожалению, в работах [37, 38], которые описывают системы обработки текстов на русском языке, не приводятся значения метрики ROUGE, поэтому нет возможности сравнить эти результаты с нашими. Также мы пришли к выводу, что некорректно сравнивать результаты работы нашей системы с результатами работы систем для английского языка, такими как, например, [90], т.к. низкие значения ROUGE могут быть связаны с особенностями языкового строя. В частности, русский язык является флективным языком с развитой морфологией, к тому же порядок слов в русском языке относительно свободный.

4.2.2 Метрика RAV

Для оценки точности предложенного метода автореферирования (описанного в главе 3) вводится RA-значение (релевантность аннотации):

$$RAV = |Count_1 - Count_2|.$$

Значения $Count_1$ и $Count_2$ рассчитываются следующим образом.

$$Count_1 = \sum_{i=1}^N \alpha_i \cdot f(w_i),$$

где N – количество ключевых слов, идентифицированных системой. Еще одна особенность состоит в том, что ключевые слова здесь – это не фразы, и не слова в их нормальной форме, а стеммы слов (без окончаний).

α_i – количество раз, когда ключевое слово появилось в автореферате;

$f(w_i)$ – частота появления ключевого слова в тексте статьи ($0 < f(w_i) < 1$).

$$Count_2 = \sum_{i=1}^N \alpha'_i \cdot f(w_i), \text{ где}$$

α'_i – это количество раз, когда ключевое слово появилось в полученном абзаце.

Будем считать, что авторская аннотация и автоматически полученный реферат схожи, если выполняется следующее условие:

$$|Count_1 - Count_2| < h, \text{ где}$$

Пороговое значение h подбирается экспертом эмпирически.

В таблице 15 изображен фрагмент результатов сравнения двух аннотаций: авторской и автоматически сгенерированной системой.

Таблица 15 – Оценка результатов

RAV	Авторская аннотация ($Count_1$)	Автоматически сгенерированная аннотация ($Count_2$)
0.0208	0.0229	0.0437
0.0246	0.0238	0.0484
0.0661	0.1057	0.0396
0.0106	0.022	0.0114
0.0121	0.0073	0.0194
0.0343	0.0453	0.011
0.0017	0.0313	0.033
0.0273	0.0322	0.0595
0.0525	0.1264	0.0739
0.0036	0.3276	0.324

В результате проведенных экспериментов можно определить, что при среднем значении $h = 0,0254$ автоматически сгенерированные и авторские рефераты наиболее схожи.

Предварительные эксперименты показали, что точность системы составляет около 60 %. Однако, следует отметить, что для оценки корректности метрики RAV необходимо провести дополнительные эксперименты.

4.2.3 Экспертная оценка

Точность полученных аннотаций, оцененная экспертами, оказалась значительно выше. Экспертная оценка результатов реферирования показала, что 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них (что, на самом деле, не всегда свидетельствует о плохом качестве реферата), и только 13,57 % представляли собой некорректно отобранные фрагменты текстов. Следует заметить, что полученная нами экспертная оценка выше, чем в работах [37, 38].

Нами было замечено, что авторы часто используют синонимы, перефразируют и меняют местами предложения. Экспертная оценка подтверждает, что порядок предложений в аннотации часто не влияет на ее общий смысл. Однако метрика ROUGE не учитывает это. Кроме того, иногда автоматически сформированная аннотация получается длиннее, чем хотелось бы (около 500 слов вместо 250). Это связано со стилем изложения самой статьи, и чаще всего означает, что в тексте имеется много содержательных предложений.

4.2.4 Точность, полнота, F-мера

Дополнительно мы рассмотрели еще один способ оценки полученных результатов при помощи точности, полноты и F-меры, которые вычисляются способом, похожим на [29, 36]. Поясним подробнее. Предположим, что автоматически полученная аннотация содержит в себе

множество W_1 ключевых слов и многословных терминов, множество V_1 специальных слов из научных и технических текстов, множество дискурсивных маркеров и коннекторов D_1 . Объединение этих множеств обозначим $N_1: N_1 = W_1 \cup V_1 \cup D_1$. Аналогичные множества можно выделить в эталонной авторской аннотации $N_2: N_2 = W_2 \cup V_2 \cup D_2$. Тогда точность, полноту и F-меру будем вычислять по следующим формулам:

$$Precision = \frac{|N_1 \cap N_2|}{|N_1|}, \quad Recall = \frac{|N_1 \cap N_2|}{|N_2|}, \quad F_{measure} = \frac{2 \cdot P \cdot R}{P + R}$$

Сравнительная оценка результатов автореферирования приведена в таблице 16.

Таблица 16 – Оценка результатов автореферирования

Система	Метод	Точность, %	Полнота, %	F-мера, %
Marcu (1998)	комбинация риторического анализа, структурного и позиционного методов	73.53	67.57	70.42
Trevgoda (2009)	риторический анализ	67.03	64.81	66.03
t-conspectus (2016)	TF-IDF	10	22.2	36.3
Open Text Summarizer (2016)	статистический	12	24.2	38.5
Scientific Text Summarizer (2018)	предложен в данной работе	75.23	68.21	71.55

Преимущество предложенных формул состоит в том, что они позволяют оценить вклад каждого из признаков и разных комбинаций этих признаков в общую оценку результата. Например, можно оценить вклад только маркеров и коннекторов, или только специальной научной лексики, или и того, и другого, но без ключевых слов и выражений и т.д. Возможное улучшение предложенного в данной статье алгоритма, по нашему мнению, состоит в том, чтобы дополнить правила удаления менее важных предложений, увеличить количество шаблонов для сглаживания, расширить списки маркеров и коннекторов.

4.3 Выводы по главе 4

Выбор базового алгоритма построения униграммных тематических моделей осуществлялся на основе эксперимента. Для его проведения была подготовлена коллекция текстов научных статей на русском языке на основе выложенных в открытом доступе архивов научных журналов. Размер коллекции составляет около 260 текстов.

Оценка результатов выполнялась при помощи следующих метрик, реализованных в библиотеке BigARTM и описанных в работе [67]: перплексия, разреженность матриц Φ и Θ ,

доля фоновых слов, мощность ядер тем, чистота ядер тем, контрастность ядер тем. В результате экспериментов по полученным значениям метрик было принято решение использовать алгоритм ARTM в реализации библиотеки BigARTM [63]. Был произведен подбор параметров алгоритма для достижения наилучших результатов на подготовленной выборке.

Оценка качества автореферирования осуществлялась на собранной нами коллекции по находящимся в открытом доступе материалам международного научно-практического журнала «Программные продукты и системы» за 2010-2018 гг. Объем этой коллекции составил 1200 статей. Результаты оценивались несколькими различными способами.

Были получены следующие значения: точность 75.23 %, полнота 68.21 % и F-мера 71.55 %, что подтверждает эффективность предложенных методов. Экспертная оценка качества результатов автореферирования показала, что 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них (что, на самом деле, не всегда свидетельствует о плохом качестве реферата), 13,57 % представляли собой некорректно отобранные фрагменты текстов. В Приложении В приведены промежуточные и конечные результаты работы системы.

Нами было замечено, что ключевые слова, заданные авторами, могут вообще отсутствовать в тексте или присутствовать в виде аббревиатур, в то время как непосредственно в тексте подобные термины используются в полной записи. В аннотациях авторы часто используют синонимы, перефразируют и меняют местами предложения. Экспертная оценка подтверждает, что порядок предложений в аннотации часто не влияет на ее общий смысл. Кроме того, иногда автоматически сформированная аннотация получается длиннее, чем хотелось бы (около 500 слов вместо 250). Это связано с авторским стилем изложения, и чаще всего означает, что в тексте имеется много содержательных предложений.

Возможное улучшение существующего алгоритма в таких ситуациях состоит в том, чтобы добавить дополнительные правила удаления менее важных предложений. Кроме того, планируется увеличить количество шаблонов для сглаживания, дополнить списки маркеров и коннекторов.

Заключение

В работе предложен оригинальный подход к автоматическому реферированию научно-технических текстов с использованием методов тематического моделирования и риторического анализа. Процесс формирования реферата состоит из шести основных этапов: предварительная обработка текста; построение тематических моделей (униграммной и расширенной); риторический анализ и формирование квазиреферата; оценка весов предложений; выбор наиболее важных предложений; сглаживание полученного текста аннотации. Система реализована на языке Python3, также используется инструмент для работы с базами данных PostgreSQL. Используются внешние библиотеки Scikitlearn, Gensim, TensorFlow, NLTK, BigARTM, Flask и некоторые другие. Для формального описания преобразований текста применяются формулы исчисления предикатов. Для построения униграммных тематических моделей используется алгоритм ARTM в реализации библиотеки BigARTM. Расширение униграммной модели многословными терминами осуществляется при помощи модификация алгоритма RAKE, который был адаптирован для работы с текстами на русском языке.

Основные результаты исследования:

1. Разработан гибридный метод, который позволяет получать рефераты (аннотации) высокого качества и определять темы текстов в виде набора ключевых терминов. Предложенный метод основан на использовании лингвистической базы знаний, графовом представлении текстов и машинном обучении.
2. Формально описана методика обнаружения важных элементов в тексте, базирующаяся на понятиях теории риторических структур. Создана лингвистическая база данных на основе анализа подязыка рефератов, используемая для оценки весов предложений квазиреферата.
3. Предложен алгоритм построения расширенных тематических моделей коллекций текстовых документов.
4. Описана процедура сглаживания предложений, позволяющая сделать текст полученного реферата (аннотации) более связным и последовательным.
5. Предложенные модели, методы и алгоритмы реализованы в виде системы, позволяющей автоматически формировать аннотации статей научно-технической тематики.
6. Собрана коллекция текстов научных статей на русском языке (около 1200 текстов) для проведения экспериментов. Проведены вычислительные эксперименты, подтверждающие эффективность предложенных методов и алгоритмов.

В дальнейшем планируется провести тестирование на научных текстах на казахском языке. Для улучшения полученных оценок будет увеличено количество шаблонов для сглаживания, дополнены списки маркеров и коннекторов. Будут проведены эксперименты с текстами из различных научных областей.

Список сокращений и условных обозначений

ARTM – Additive Regularization for Topic Modeling

BigARTM – открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределённая реализация методов вероятностного тематического моделирования на основе аддитивной регуляризации.

DLCS – Direct Lexical Chain Score

DLCSS – Direct Lexical Chain Span Score

DT – Decision Trees

GPR – Google’s Pagerank

HITS – Hyperlinked Induced Topic Search

HMM – Hidden Markov Model

IF-IDF – Term Frequency-inverse document frequency

LC – Lexical Chain

LCS – Lexical Chain Score

LCSS – Lexical Chain Span Score

LSA – Latent Semantic Analysis

ME – Maximum Entropy

MLSA – Meta Latent Semantic Analysis

NB – Naïve Bayes

NMF – Non-Negative Matrix Factorization

NN – Neural Networks

POK – Position of a Keyword

RST – Rhetorical Structure Theory

SDD – Semi-Discrete Decomposition

SLSS – Sentence Level Semantic Analysis

SNMF – Symmetric Nonnegative Matrix Factorization

SVD – Singular Value Decomposition

SVM – Support Vector Machine

TF – Term Frequency

Автоаннотирование – извлечение наиболее важных сведений из одного или нескольких документов и составление их краткого описания.

Автореферирование – это составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных отчетов.

Аннотация – краткое изложение содержания документа, дающее общее представление о его теме, т.е. в отличие от реферата выполняющее лишь сигнальную функцию (есть публикация на определенную тему).

АРТМ – аддитивная регуляризация тематических моделей.

Квазиреферат – перечень наиболее информативных предложений текста.

Ключевое предложение – предложение, которое содержит несколько (два и более) ключевых слов.

Ключевое слово – слово, относящееся к основному содержанию текста и позволяющее выявить его тематику.

Ключевое словосочетание – сочетание слов, среди которых есть одно или несколько ключевых.

Коннекторы – группы слов, заменяющие маркеры и характеризующие определенные риторические отношения.

Маркеры (дискурсивные маркеры) – это слова или фразы, которые не имеют реального лексического значения, но вместо этого обладают важной функцией формировать разговорную структуру, передавая намерения говорящих при разговоре.

Многословное выражение (Multiword Extraction, MWE) – устойчивая последовательность слов (n-грамма), имеющая определенную семантику в контексте заданной предметной области и обладающая значительной частотой встречаемости по сравнению с другими n-граммами.

Реферат – связный текст, который кратко выражает центральную тему, предмет.

Тема – набор терминов (слов и словосочетаний), характеризующих принадлежность текста к определенной области знаний.

Тематическое моделирование – способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.

ТРС – теория риторических структур

УДК – универсальная десятичная классификация

ЭДЕ – элементарная дискурсивная единица

Литература

1. *Луканин А.В.* Автоматическая обработка естественного языка. Челябинск: Изд. центр ЮУрГУ, 2011. 70 с.
2. *Bharti S.K., Babu K.S., Jena S.K.* Automatic Keyword Extraction for Text Summarization: A Survey. 2017. [Electronic resource] URL: <https://arxiv.org/ftp/arxiv/papers/1704/1704.03242.pdf> (дата обращения 11.10.2018)
3. *Ступин В.С.* Система автоматического реферирования методом симметричного реферирования // Компьютерная лингвистика и интеллектуальные технологии. Труды межд. Конференции «Диалог 2004». М.: Наука, 2004. С. 579-591.
4. *Kupiec J., Pederson J. and Chen F.* A trainable document summarizer.// In Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval, Seattle, 1995. pp. 68-73.
5. *Танатар Н.В., Федорчук А.Г.* Интеллектуальные поисково-аналитические системы мониторинга СМИ // Научно-практический и теоретический сборник. Киев, 2008. 477 с.
6. *Михаилин А.* Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах. 2000. [Электрон. ресурс] URL: <http://citforum.ru/programming/digest/avtestlang.shtml> (дата обращения: 11.10.2018)
7. *Харламов А.А.* Автоматический структурный анализ текстов // Открытые системы. Москва. 2002. №10. С.16-22.
8. *Кутукова Е.С.* Технология Text mining // SWorld: Перспективные инновации в науке, образовании, производстве и транспорте. Одесса, 2013. с.136-138.
9. RCO Fact Extractor Desktop. 2000. [Электрон. ресурс] URL: http://www.rco.ru/?page_id=4875 (дата обращения: 11.10.2018)
10. *Бурмистров А.С., Свиридова О.В.* Экспертная оценка программных продуктов для аннотирования документов // Постулат. 2017. № 5. [Электрон. ресурс] URL: <http://e-postulat.ru/index.php/Postulat/article/viewFile/567/588> (дата обращения: 06.12.2018)
11. *Фисун А.П., Еременко В.Т., Минаев В.А., Зернов В.А., Константинов И.С., Коськин А.В., Белевская Ю.А., Дворянкин С.В.* Организационные и технико-экономические основы: учебник для вузов. Орел: ОрелГТУ, ОГУ, 2009. 171 с.
12. *Luhn H.* The automatic creation of literature abstracts // In IBM Journal of Research and Development, New York, 1958. Vol. 2(2). P. 159–165.
13. *Лукашевич Н.В.* Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. М.: 1997 С. 188–191.

14. *Лукашевич Н.В., Добров Б.В.* Построение структурной тематической аннотации текста // Труды международного семинара Диалог-98, Т. 2. М. 1998. С. 795–802.
15. *Лукашевич Н., Добров Б.* Автоматическое аннотирование новостного кластера на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции "Диалог" 2009. Т. 8. Изд-во РГГУ Москва, 2009. С. 27–31.
16. *Яцко В.А.* Симметричное реферирование: теоретические основы и методика // НТИ. Серия 2. Информационные процессы и системы. 2002. № 5. С. 18-28.
17. *Вичева О.Н.* Подходы к автоматическому обзорному реферированию группы текстов одной тематики // Проблемы современной прикладной лингвистики: сб. науч. статей. Минск: МГЛУ, 2014. С. 246-252.
18. *Edmundson H.P.* New methods in automatic extracting // Journal of the ACM (JACM). 1969. V.16. №2. P. 264–285.
19. *Бумаков А.* T-CONSPECTUS. URL: <http://tconspectus.pythonanywhere.com/about#algorithm>
20. *Andonov F., Slavova V., Petrov G.* On the Open Text Summarizer // International Journal "Information Content and Processing". Vol. 3. N 3. 2016. URL: <http://www.foibg.com/ijicp/vol03/ijicp03-03-p05.pdf>
21. Automatic Text Summarization Using Latent Semantic Analysis // Programming and Computer Software. 2011. V. 37. № 6. P. 299–305.
22. *Babar S.A., Pallavi D. Patil.* Improving Performance of Text Summarization // Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India. Amsterdam, Elsevier, 2015. P. 354–363.
23. *Wang Y. A., Jun Ma.* Comprehensive Method for Text Summarization Based on Latent Semantic Analysis // Proceedings of Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013. Berlin, Springer Berlin Heidelberg, 2013. P. 394–401.
24. *Kupiec J., Pedersen J., Chen F.* A Trainable Document Summarizer // Proceeding SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval Seattle, WA, USA. 1995. P. 68–73.
25. *Kumar M., Das D., Agarwal S., Rudnicky A.* Non-textual event summarization by applying machine learning to template-based language generation // Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009. Suntec, 2009. P. 67–71.
26. *Saggion H.* A classification algorithm for predicting the structure of summaries // Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009. Suntec, 2009. P. 31–38.

27. *Maâloul M. H.* Approche hybride pour le résumé automatique de textes. Application à la langue arabe // Theses. Université de Provence – Aix-Marseille I, 2012. Français. [Electronic resource] URL: <https://tel.archives-ouvertes.fr/tel-00756111/> (дата обращения: 11.10.2018).
28. *Mann W. C., Thompson S. A.* Rhetorical structure theory: Toward a functional theory of text organization // *Interdisciplinary Journal for the Study of Discourse*. 1988. V. 8, № 3. P. 243–281.
29. *Ono K., Sumita K., Miike S.* Abstract generation based on rhetorical structure extraction // *Proceedings of Coling '94*. Morristown, NJ, USA. 1994. P. 344–348.
30. *Marcu D.* Improving summarization through rhetorical parsing tuning // *Proceedings of The Sixth Workshop on Very Large Corpora*. Montreal, Canada. 1998. P. 206–215.
31. *Strzalkowski T., Stein G., Wang J., Wise B.* A Robust Practical Text Summarizer // *Advances in Automatic Text Summarization*. Cambridge, Massachusetts, MIT Press, 1999. P. 137–154.
32. *Ананьева М.И.* Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур / М.И. Ананьева, М.В. Кобозева // Тр. Междунар. конф. «Диалог», 2016. [Электрон. ресурс] URL: www.dialog-21.ru/media/3460/ananyeva.pdf (дата обращения: 11.10.2018).
33. *Teufel S., Moens M.* Summarizing scientific articles: experiments with relevance and rhetorical status // *Computational Linguistics*. 2012. Vol. 28(4), pp. 409–445.
34. *Bosma W.* Query-Based Summarization using Rhetorical Structure Theory // *15th Meeting of CLIN*. 2015. pp. 29–44.
35. *Huspi S.H.* Improving Single Document Summarization in a Multi-Document Environment // PhD thesis. 2017. RMIT University, Melbourne, Australia, 190 p.
36. *Mithun S.* Exploiting rhetorical relations in blog summarization // PhD thesis, 2012. Concordia University, Montreal, Canada. 230 p.
37. *Тревгода С.А.* Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений, [Текст]: автореф. дис. на соиск. учен. степ. канд. тех. наук (05.13.01) / Тревгода Сергей Александрович, Санкт-Петербургский государственный электротехнический университет, Санкт-Петербург, 2009. – с.15
38. *Осминин П.Г.* Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод [Текст]: автореф. дис. на соиск. учен. степ. канд. филол. наук (10.02.21) / Осминин Павел Григорьевич; Южно-Уральский гос. унив. – Челябинск, 2016. – 239 с.
39. *Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* Towards building a discourse-annotated corpus of Russian // *Computational Linguistics and Intellectual Technologies*. 2017. Iss. 16 (23). V. 1. pp. 194–204.

40. *Khan A., Salim N., Kumar Y.* A Framework for multi-document abstractive summarization based on semantic role labelling // *Applied Soft Computing*. 2015. Vol. 30. pp. 737–747.
41. *Murray, G.* Abstractive Meeting Summarization as a Markov Decision Process // *Proceedings of 28th Canadian Conference on Artificial Intelligence, Canadian AI 2015, Halifax, Nova Scotia, Canada, June 2-5, 2015*. Switzerland, Springer International Publishing, 2015. P. 212–219.
42. *Genest P.-E., Lapalme G.* Framework for Abstractive Summarization using Text-to-Text Generation // *In Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon, USA. 2011. pp. 64–73.
43. *Lloret E., Roma-Ferri M. T., Palomar M.* COMPENDIUM: A text summarization system for generating abstracts of research papers // *Data & Knowledge Engineering*. 2013. Vol. 88. pp. 164–175.
44. *Hovy E., Lin Ch.-Y.* Automated text summarization and the SUMMARIST system // *Proceedings of the TIPSTER Text Program*. 1998. pp. 197–214.
45. *Saggion H., Lapalme G.* Generating indicative-informative summaries with SumUM // *Computational Linguistics*. 2002. V. 28. N 4. P. 497–526.
46. *Foster G. F.* Statistical lexical disambiguation: Master's thesis. 1991. 340 p.
47. *Plaza L., Diaz A., Gervas P.* Concept-graph based Biomedical Automatic Summarization using Ontologies // *Coling 2008: Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing*. Manchester, 2008. P. 53–56.
48. Unified Medical Language System (UMLS). 2016. [Electronic resource] URL: <http://www.nlm.nih.gov/research/umls/> (дата обращения: 11.10.2018)
49. *Aronson A. R.* Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program // *Proceedings of American Medical Informatics Association*. 2001. P. 17–21.
50. *Farzindar A., Lapalme G.* Legal text summarization by exploration of the thematic structures and argumentative roles // *Text Summarization Branches Out Conference, ACL*. Barcelona, Spain. 2004 P. 27–38.
51. *Galgani F., Compton P., Hoffmann A.* Combining Different Summarization Techniques for Legal Text // *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012*. Avignon, France. 2012. P. 115–123.
52. *Megala S., Kavitha A., Marimuthu A.* Feature Extraction Based Legal Document Summarization // *International Journal of Advance Research in Computer Science and Management Studies*. 2014. V.2. Issue 12. P. 346–352.

53. *Lloret E., Boldrini E., Vodolazova T., Martínez-Barco P., Muñoz R., Palomar M.* A novel concept-level approach for ultra-concise opinion summarization // *Expert Systems with Applications*. 2015. Vol. 42, Issue 20. P. 7148–7156.
54. *Brügmann S., Bouayad-Aghab N., Burga A., Carrascosa S., Ciaramella A., Ciaramella M., Codina-Filba J., Escorsa E., Judea A., Mille S., Müller A., Saggion H., Ziering P., Schütze H., Wanner L.* Towards content-oriented patent document processing: Intelligent patent analysis and summarization // *World Patent Information*. 2015. Vol. 40. P. 30–42.
55. *Mahdabi P., Andersson L., Hanbury A., Crestani F.* Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization. 2011. [Electronic resource] URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.7897&rep=rep1&type=pdf> (дата обращения: 11.10.2018).
56. *Wanner L.* Generation of Patent Abstracts: A Challenge for Automatic Text Summarization // *Proceedings of the SEPLN 2012 workshops: E-LKR and ATSF*. 2012. [Electronic resource] URL: http://ceur-ws.org/Vol-882/elkr_atsf_2012_keynote.pdf (дата обращения: 11.10.2018).
57. *Chieze E.* An Automatic System for Summarization and Information Extraction of Legal Information // *Semantic Processing of Legal Texts / Enrico Francesconi, Simonetta Montemagni, Wim Peters, Daniela Tiscornia*. Heidelberg, 2010. pp. 216-234.
58. *Goldstein A.* Generation of Natural-Language Textual Summaries from Longitudinal Clinical Records // *Studies in Health Technology and Informatics*. 2015. 216. P. 594–598.
59. *Goldstein A.* An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data // *Journal of Biomedical Informatics*. 2016. Vol. 61. P. 159–175.
60. *Анисимов А.В., Марченко А. А.* Ассоциативное реферирование естественно-языковых текстов // *Штучний інтелект*. 2006. № 3. С. 488–492.
61. *Попов М.Ю., Заболева-Зотова А.В., Фоменков С.А.* Визуализация семантической структуры и реферирование текстов на естественном языке. 2003. [Электрон. ресурс] URL: <http://www.dialog-21.ru/media/2725/popov.pdf> (дата обращения: 11.10.2018).
62. *Коришунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // *Труды Института системного программирования РАН*. 2012. С. 215–242.
63. *Воронцов К.В., Фрей А.И., Апишев М.А., Ромов П.А., Янина А.О., Суворова М.А.* BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // *Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL'2015, Обнинск, 13-16 октября 2015.* [Электрон. ресурс] URL: <http://www.machinelearning.ru/wiki/images/e/e4/Voron15damdid.pdf> (дата обращения: 06.12.2018)

64. Батура Т. В., Стрекалова С. Е. Подход к построению расширенных тематических моделей текстов на русском языке // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 5–18.
65. Hofmann T. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99). 1999. pp. 289–296.
66. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. N 3. pp. 993–1022.
67. Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.) Вып. 13(20). М.: Изд-во РГГУ, 2014. С. 676–687.
68. Купяткова И.С., Карпов А.А. Аналитический обзор систем распознавания русской речи с большим словарем // Труды СПИИРАН, 2010, Т. 12, с. 7–20.
69. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». М.: РГГУ, 2007 с.70-75.
70. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Machine Learning; Models, Technologies and Applications (MLMTA), LasVegas. 2003. pp. 273-280.
71. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // Text Mining: Applications and Theory. 2010. pp. 3-20.
72. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. 2014. 513 p.
73. Gülşen Eryiğit, Joakim Nivre, Kemal Oflazer. Dependency Parsing of Turkish // Computational Linguistics. 2008. Vol. 34. No. 3. P. 357–389.
74. Kemal Oflazer. Two-level Description of Turkish Morphology // Literary and Linguistic Computing. 1994. Vol. 9. No. 2. P. 137–148.
75. Жуманов Ж. М. Разработка грамматики связи для синтаксического анализа казахского языка // Вестн. КазНУ. Серия: Математика, механика, информатика. 2012. № 2 (73). С. 71–80.
76. Тулеев У. А., Жуманов Ж. М., Рахимова Д. Р. Моделирование семантических ситуаций времен казахского языка при машинном переводе // Вестн. КазНУ. Серия: Математика, механика, информатика. 2012. № 4 (75). С. 99–107.

77. Белоногов Г. Г., Зеленков Ю. Г. Алгоритм автоматического анализа русских слов // Вопросы информационной теории и практики. 1985. № 53. С. 62–93.
78. Porter M. F. An algorithm for suffix stripping // Program: Electronic Library and Information Systems. 1980. Vol. 14. № 3. pp. 130–137.
79. Temperley D. An Introduction to the Link Grammar Parser. 2014. [Electronic resource] URL: <http://www.abisource.com/projects/link-grammar/dict/introduction.html#1> (дата обращения 06.12.2018)
80. Kessikbayeva G., Cicekli I. Rule Based Morphological Analyzer of Kazakh Language // Proc. of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. 2014. P. 46–54.
81. Özlem İstek. A Link Grammar for Turkish. Thesis. Ankara: Bilkent University, 2006. 135 p.
82. Куликовская Л. К., Мусаева Э. Н. Грамматика казахского языка в таблицах и схемах в сопоставлении с грамматикой русского языка. Алмата, 2006. 76 с.
83. Сусов А. А. Моделирование дискурса в терминах теории риторической структуры // Вестник Воронежского государственного университета. Серия: Филология. Журналистика. 2006. №2. С. 133–138.
84. Баранов А. Г. Функционально-прагматическая концепция текста. Ростов н/Д : изд. Рост. ун-та, 1993. 182 с.
85. Fraser B. What are discourse markers? // Journal of pragmatics. 1999. Vol. 31. No. 7. P. 931–952.
86. Палатовская Е. В. Дискурсивный анализ и теория риторической структуры // Науковий вісник кафедри ЮНЕСКО КНЛУ. Сер. Філологія. Педагогіка. Психологія. 2014. Вип. 29. С. 89–95.
87. Louwerse M. An Analytic and Cognitive Parameterization of Coherence Relations. Cambridge, 2001. 320 p.
88. Das D., Martins A. A. Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course at CMU. 2007. pp. 192–195.
89. Lin Ch.Y. ROUGE: A Package for Automatic Evaluation of Summaries // Workshop On Text Summarization Branches Out. 2004. pp. 74–81.
90. Zhang J.J., Chan H.Y., Fung P. Improving lecture speech summarization using rhetorical information // 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). 2007. pp. 195–200.

Публикации автора

91. Батура Т.В., Бакиева А.М. Создание системы автоматического реферирования научных текстов // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 74–86.

92. Бакиева А.М., Батура Т.В. Исследование применимости теории риторических структур для автоматической обработки научно-технических текстов // *Cloud of Science*. 2017. Т. 4. № 3. С. 450–464.
93. Бакиева А.М., Батура Т.В., Еримбетова А.С., Митьковская М.В., Семенова Н.А. Исследование грамматики связей на примере казахского и турецкого языков // *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии*. 2016. Т. 14, № 3. С. 5–14.
94. Баракхнин В.Б., Бакиева А.М., Бакиев М.Н., Тажиббаева С.Ж., Батура Т.В., Лукпанова Л.Х. Стемматизация и генерация словоформ в казахском языке для систем автоматической обработки текстов // *Вычислительные технологии*. Новосибирск: ИВТ. 2017. Т. 22, № 4. С. 11–21.
95. Баракхнин В.Б., Федотов А.М., Бакиева А.М., Бакиев М.Н., Тажиббаева С.Ж., Батура Т.В., Кожемякина О.Ю., Тусупов Д.А., Самбетбаева М.А., Лукпанова Л.Х. Алгоритмы генерации и стемматизации словоформ казахского языка // *Cloud of Science*. 2017. Т. 4. № 3. С. 434–449.
96. Федотов А.М., Тусупов Д.А., Самбетбаева М.А., Бакиева А.М., Еримбетова А.С., Идрисова А.И. Модель определения нормальной формы слова для казахского языка // *Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии*. 2015. Т. 13, № 1. С. 107–116.
97. Batura T.V., Murzin F.A., Semich D.F., Sagnayeva S.K., Tazhibayeva S.Zh., Bakiyev M.N., Yerimbetova A.S., Bakiyeva A.M. Using the Link grammar parser in the study of Turkic languages // *Eurasian journal of mathematical and computer applications*. ISSN: 23066172. Astana: L.N. Gumilyov Eurasian National University, 2016. V. 4. Iss. 2. pp. 14–22.
98. Yerimbetova A.S., Murzin F.A., Batura T.V., Sagnayeva S.K., Semich D.F., Bakiyeva A.M. Estimation of the degree of similarity of sentences in a natural language based on using the Link Grammar Parser program system // *Journal of Theoretical and Applied Information Technology*, 2016. Vol. 86. N. 1. P. 68–77.
99. Yerimbetova A.S., Murzin F.A., Batura T.V., Sagnayeva S.K., Tazhibayeva S.Zh., Bakiyeva A.M. Link Grammar Parser for Turkic Languages and algorithms for estimation the relevance of documents // *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT-2016)*. 12-14 October 2016, Baku, Azerbaijan. 2016. pp. 104-107.
100. Barakhnin V.B., Bakiyeva A.M., Fedotov A.M., Bakiyev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tussupov D.A., Sambetbaiyeva M.A., Lukpanova L.Kh. The software system for the study the morphology of the Kazakh language // *The European Proceedings of Social and Behavioural Sciences*. 2017. V. XXXIII. P. 18-27.

101. *Бакиева А.М., Батура Т.В.* Свидетельство Роспатента о государственной регистрации программы для ЭВМ Система автоматического реферирования и определения тем научных текстов «Scientific Text Summarizer» № 2018661835 от 19.09.2018.
102. *Бакиева А.М.* Свидетельство Роспатента о государственной регистрации программы для ЭВМ Морфологическая система «Стемматизация и генерация словоформ казахского языка» № 2018614456 от 19.12.2017.
103. *Еримбетова А.С., Батура Т.В., Мурзин Ф.А., Сагнаева С.К., Бакиева А.М.* Свидетельство о государственной регистрации прав на объект авторского права Министерства Юстиции Республики Казахстан «Қазақ және түрік тілдеріне арналған LINK GRAMMAR PARSER синтаксистік талдағышы» запись в реестре № 743 от 17.04.2017 г.
104. *Batura T.V., Murzin F.A., Bakiyeva A.M., Yerimbetova A.S.* The methods of estimation of the degree of similarity of sentences in a natural language based on the link grammar // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2014. Is. 37. P. 55–69. URL: http://bulletin.iis.nsk.su/files/article/batura_v8.pdf
105. *Batura T.V., Murzin F.A., Semich D.F., Bakiyeva A.M., Yerimbetova A.S.* On some graphs connected with texts in a natural language, link grammar and the summarization process // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2015. Is. 38. p. 37–49.
106. *Бакиева А.М., Батура Т.В., Федотов А.М.* Методы и системы автоматического реферирования текста // Вычислительные технологии. 2015. Т. 20, № 3. С. 263–274.
107. *Мурзин Ф.А., Батура Т.В., Бакиева А.М., Еримбетова А.С.* Методы определения степени близости предложений на естественном языке на основе грамматики связей // Наука и мир. Волгоград: Научное обозрение, 2015. № 3 (19). Т. 2. С. 61–67.
108. *Бакиева А.М.* Подходы к созданию моделей определения тем текстов на тюркских языках // Труды XVI Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям (УМ-2015). 2015. Красноярск, Россия, 28-30 октября 2015. С. 60.
109. *Bakiyeva A.M., Batura T.V., Yerimbetova A.S., Mit'kovskaya M.V., Semenova N.A.* Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2016. Is. 40. pp. 37–51. URL: http://bulletin.iis.nsk.su/files/article/batura_3.pdf
110. *Murzin F.A., Batura T.V., Semich D.F., Sagnayeva S.K., Bakiyeva A.M., Yerimbetova A.S., Mit'kovskaya M.V., Semenova N.A.* Research of link grammar for kazakh and turkish languages // Вестник КазНУ. Алматы, 2016. № 4 (116). С. 684–691.

111. *Бакиева А.М.* Методы автоматического анализа текстов на казахском языке // Материалы 55-й Международной научной студенческой конференции, МНСК – 2017, г. Новосибирск, 17 - 20 апреля 2017 г., С. 151.
112. *Бакиева А.М., Еримбетова А.С.* Исследование грамматики связей на примере турецкого и казахского языка // Материалы 54-й Международной научной студенческой конференции, МНСК – 2016, г. Новосибирск, 16 - 20 апреля 2016 г, С. 163.
113. *Еримбетова А.С., Бакиева А.М.* Модели определения релевантности текста и задача реферирования // Материалы 54-й Международной научной студенческой конференции, МНСК – 2016, г. Новосибирск, 16 - 20 апреля 2016 г, С. 167.
114. *Бакиева А.М.* Стемматизация и генерация словоформ казахского языка для систем автоматической обработки текстов // XVII Всероссийская конференция молодых учёных по математическому моделированию и информационным технологиям УМ-2016, г. Новосибирск, 30 октября - 3 ноября 2016 г., С. 63.
115. *Barakhnin V., Bakiyeva A.M., Batura T.* Stemming and word forms generation in automatic text processing systems in the Kazakh language // The 15th International Scientific Conference «Information Technologies and Management». Theses. Riga, April 28-29, 2017. Riga: ISMA University, 2017. ISSN 1691-2489. P. 85-86. URL: http://isma.lv/FILES/SCIENCE/IT&M2017_THESES/02_CMIT/21_IT&M2017_Barakhnin.pdf
116. *Батура Т.В., Бакиева А.М.,* Применение теории риторических структур для автоматической обработки текстов // Марчуковские научные чтения - 2017 (MSR 2017) // Новосибирск: Омега Принт, 2017. Новосибирск, 25 июня–14 июля 2017 г. С. 149.
117. *Batura T.V., Murzin F.A., Semich D.F., Yerimbetova A.S., Bakiyeva A.M.* Link Grammar Parser and estimation of the document relevance to the search query // Марчуковские научные чтения - 2017 (MSR 2017). Тезисы. Новосибирск: Омега Принт, 2017. Новосибирск, 25 июня–14 июля 2017 г. С. 200.
118. *Мурзин Ф.А., Еримбетова А.С., Сагнаева С.К., Батура Т.В., Бакиева А.М., Семич Д.Ф.* Алгоритмы и программные инструменты для определения релевантности текста поисковому запросу и определения тем текстов // Труды Международной конференции «Актуальные проблемы чистой и прикладной математики». Алматы: ИМиММ, 2017. Алматы, 22–25 августа 2017 г. С. 141–142.
119. *Мурзин Ф.А., Еримбетова А.С., Батура Т.В., Бакиева А.М., Семич Д.Ф., Ефимова Л.В.* О новых инструментах поиска информации на основе грамматики связей // Интеллектуальный анализ сигналов, данных и знаний: методы и средства. Сборник

статей Всероссийской научно-практической конференции с международным участием. Новосибирск: НГТУ, 2017. С. 161–166.

120. *Бакиева А.М., Батура Т.В.* Система автоматического реферирования и определения тем научно-технических текстов // XVI Российская конференция «Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4–7 декабря 2017 г.). 2017. Новосибирск: ИВТ СО РАН. С. 75–80.
121. *Мурзин Ф.А., Батура Т.В., Еримбетова А.С., Бакиева А.М., Семич Д.Ф., Ефимова Л.В.* О системе поиска информации на основе грамматики связей // XVI Российская конференция «Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4–7 декабря 2017 г.). 2017. Новосибирск: ИВТ СО РАН. С. 100–114.
122. *Бакиева А.М., Батура Т.В.* Методы автоматического реферирования и определения тем текстов // Материалы XVIII Всероссийской конференции молодых ученых по математическому моделированию и информационным технологиям. г. Иркутск, Россия, 21-25 августа 2017. Новосибирск: ИВТ СО РАН, 2017. С. 65.
123. *Батура Т. В., Бакиева А.М.* Применение теории риторических структур в системах автоматической обработки текстов // TurkLang – 2017. 18–21 октября, 2017. Казань. С. 18-30.
124. *Barakhnin V.B., Kozhemyakina O.Yu., Bakiyeva A.M., Sodboev M.K.* The algorithms for complex analysis of the corpuses of poetic texts in the Kazakh language // Journal of Physics: Conf. Series. 2018. V. 1117. pp. 1-7. URL: <http://dx.doi.org/10.1088/1742-6596/1117/1/012003>
125. *Баракнин В.Б., Кожемякина О.Ю., Бакиева А.М., Содбоев М.К.* Алгоритмы автоматизированной обработки поэтических текстов на казахском языке // II Международная научная конференция «Информатика и прикладная математика». 26-29 сентября 2018 года. Алматы. С. 55-64.

Приложения

Приложение А. Таблицы маркеров и коннекторов

Таблица 17 – Некоторые риторические маркеры

Elaboration(Детализация) <i>El1</i> : Вследствие (того, чего), <i>El2</i> : Кроме того, <i>EL3</i> : Например, <i>EL4</i> : в том числе, <i>El5</i> : В частности <i>El6</i> : Их можно условно разбить\разделить	Concession (Уступка) <i>Conc1</i> : Поскольку, <i>Conc2</i> : Исходя из этого, <i>Conc3</i> : Хотя,
Restatement (Переформулировка) <i>Res1</i> : То есть, <i>Res2</i> : Иными словами, <i>Res3</i> : Иначе говоря <i>Res3</i> : Между тем, практика показывает	Contrast (Контраст) <i>Cont1</i> : Однако, <i>Cont2</i> : Несмотря на то, что <i>Cont3</i> : Обратите внимание <i>Cont4</i> : Но
Purpose (Цель) <i>Pur1</i> : Для того, чтобы <i>Pur2</i> : Чтобы <i>Pur3</i> : целью которого <i>Pur4</i> : целью чего	Evidence (Обоснование) <i>Ev1</i> : Очевидно, что <i>Ev2</i> : Доказательством тому, <i>Ev3</i> : Доказательство чего, <i>Ev4</i> : Таким образом, <i>Ev5</i> : Безусловно, <i>Ev6</i> : Можно сделать вывод <i>Ev7</i> : Как показала (практика), <i>Ev8</i> : Основной же проблемой <i>Ev9</i> : Важной составляющей <i>Ev10</i> : Преимуществами системы <i>Ev11</i> : В трактовке <i>Ev12</i> : Исследования показали <i>Ev13</i> : Основной идеей <i>Ev14</i> : В настоящей работе <i>Ev15</i> : В данной работе <i>Ev16</i> : В данной статье <i>Ev17</i> : На практике приведенный алгоритм <i>Ev18</i> : Важное преимущество
Cause-Effect (Причина) <i>Cef1</i> : Почему <i>Cef2</i> : Из-за <i>Cef3</i> : Так как <i>Cef4</i> : Поэтому <i>Cef5</i> : Потому	Background (Фон) <i>Bg1</i> : При этом <i>Bg2</i> : При том <i>Bg3</i> : Для его внедрения/использования

Таблица 18 – Краткая таблица РСТ маркеров и действий

№	Название маркера	Маркер	Действие
1.	Evidence	Таким образом	mdelete_save
2.	Evidence	Очевидно, что	mdelete_save
3.	Purpose	Для того, чтобы	delete_save
4.	Purpose	Чтобы	save

5.	Restatement	То есть	save_delete
6.	Restatement	Иными словами	save_delete
7.	Elaboration	например, о том	delete_example
8.	Restatement	Иначе говоря	save_delete
9.	Restatement	Речь идет	save_delete
10.	Restatement	Другими словами	save_delete
11.	Restatement	Соответственно	save_delete
12.	Cause-Effect	Почему	Mdelete_save
13.	Cause-Effect	Из-за	Mdelete_save
14.	Cause-Effect	Так как	Mdelete_save
15.	Cause-Effect	Поэтому	save_save
16.	Cause-Effect	Потому	Mdelete_save
17.	Comparison	Больше чем,	save_delete
18.	Concession	Поскольку	delete_save
19.	Concession	Исходя из этого	mdelete_save
20.	Concession	Хотя	Mdelete_save
21.	Elaboration	Например	save_delete
22.	Elaboration	Вследствие того	save_delete
23.	Elaboration	Вследствие чего	save_delete
24.	Elaboration	Кроме того	save_delete
25.	Elaboration	В том числе	save_delete
26.	Elaboration	В частности	save_delete
27.	Elaboration	например	delete_example
28.	Elaboration	кроме того	save_mdelete
29.	Elaboration	к примеру	save_mdelete
30.	Contrast	Однако	save_save
31.	Contrast	Несмотря на то, что	save_delete
32.	Contrast	Обратите внимание	save_delete
33.	Contrast	однако	save
34.	Contrast	несмотря на то, что	save_mdelete
35.	Evidence	В данной работе	save
36.	Evidence	В статье	save
37.	Elaboration	Необходимо отметить	save_delete
38.	Elaboration	В связи с этим	save_delete
39.	Concession	Как видно из таблицы	delete_save
40.	Concession	Как видно из рисунка	delete_save
41.	Concession	Соответственно	delete_save
42.	Concession	Наконец	delete_save
43.	Restatement	В то же время	save_delete
44.	Elaboration	в том числе	delete_example
45.	Concession	В заключение отметим	save
46.	Elaboration	а именно	delete_example
47.	Background	При этом	save_delete
48.	Concession	В результате работы	save
49.	Cause-Effect	поэтому	mdelete_save
50.	Elaboration	Кроме этого	save_save"

Таблица 19 – Примеры РСТ маркеров и действий

№	Отношение	Маркер	Однояд./ многояд.	Действие	Примеры
1.	Детализация (Elaboration)	Вследствие (того, чего)	однояд	Save_delete	Это дает возможность использования в процессе создания нового программного комплекса фрагментов описания предметных областей, функциональных модулей, исходных данных и результатов вычислений, имеющихся в других комплексах. <i>Вследствие этого сокращаются сроки разработки прикладного программного обеспечения и проведения вычислительных экспериментов.</i>
2.		Кроме того		Save_delete	Целью введения послойной организации является разделение особей, принадлежащих разным видам, упрощение представления и анализа межвидовых взаимодействий, в том числе пищевых цепей и пирамид, и обеспечение управляемости программной системой. <i>Кроме того, в предлагаемой модели поддерживается возможность сохранения в клетке следов пребывания агентов.</i>
3.		Например,		Save_delete	В соответствии с методом по формулам (3) для каждого критерия вычисляются значения базового распределения доверия. <i>Например, для критерия $C1$ весом $w1c=0,6$ эти значения равны: $m1(S11)=0,072$, $m2(S12)=0,360$, $m1(S13)=0,144$, $m1(S14)=0,216$, $m1(\Theta)=0,208$.</i>
4.		например,			

5.		в том числе		Save	Данные модели предложены для построения трехуровневой системы импульсной взрывопожарной защиты любого потенциально опасного или опасного объекта, в том числе химического или нефтеперерабатывающего предприятия, атомной электростанции и т.п.
6.		В частности в частности		Save_mdelete	Если затраты на производственные ресурсы снизить достаточно сложно, то затраты на обеспечение безопасных условий труда можно значительно сократить, <u>в частности</u> , за счет проектирования новых рабочих мест, на которых исключено нарушение техники безопасности.
7.		Их можно условно разбить Их можно условно разделить Их можно разбить Их можно разделить		Save_save	Неопределенный характер температурных полей технических систем обусловлен неопределенным характером факторов, определяющих тепловой режим технической системы. <u>Их можно условно</u> разбить на три группы: факторы конструкции технической системы, факторы, возникающие при функционировании технической системы, и факторы окружающей среды.
8.	Уступка (Concession)	Поскольку		Mdelete_save	<u>Поскольку</u> встроенный тип проектов разрабатывается в рамках жестких ограничений по аппаратному ПО и пр., что соответствует особенностям разработки ПО для научной деятельности в ракетно-космической отрасли, целесообразно применить этот тип проекта.
9.		Исходя из этого		Mdelete_save	<u>Исходя из этого</u> , разработанная модель и программный комплекс могут применяться для моделирования процесса непрерывного литья цилиндрических заготовок из цветных металлов.

10.		Хотя		Save	<u>Хотя</u> использование ГЛОНАСС весьма актуально, анализ показывает, что технология позиционирования и идентификации мобильных объектов на пространственных цифровых моделях в транспортной сфере развита недостаточно.
11.	Переформулировка (Restatement)	То есть		Save_delete	Хороших результатов работы метода удалось добиться на слипшихся клетках, относящихся к разным типам. <i>То есть слипшиеся лейкоциты и эритроциты (как, например, на рисунке 4) удается корректно разделить если не всегда (2 % ошибок приходится как раз на случай слипшихся клеток), то в большинстве случаев.</i>
12.		Иными словами	однойд	Save_delete	<p>Реальные температурные поля технических систем, как показывает практика, не являются строго определенными и детерминированными, а носят неопределенный, а точнее интервальный характер. <u>Иными словами</u>, температура в каждой точке технической системы может принимать любые значения внутри некоторых интервалов своего изменения.</p> <p>Как известно, для особого семейства МЦ, называемых эргодичными, по прошествии длительного периода времени вероятность попадания случайной величины в то или иное состояние не-рестает зависеть от начального состояния цепи. <u>Иными словами</u>, при $i \rightarrow \infty P(i)ab = P(i)b$.</p> <p>Для данного случая это условие можно выразить так: возведенная в некоторую степень матрица перехода не содержит нулевых элементов. <u>Иными словами</u>, у МЦ есть вероятность через определенное число шагов перейти из любого состояния в какое-либо другое.</p>

13.		иначе говоря	двухяд	Save_delete	<p>Существующие методы моделирования температурных полей технических систем исходят из допущения, что параметры, определяющие тепловые режимы, являются детерминированными, <u>иначе говоря</u>, все данные, определяющие протекание теплового процесса и его характер в технической системе, являются полностью известными и однозначно определенными.</p> <p>Следующим модулем, в котором агрегируется совокупность действий одного уровня, последовательно реализуемых в рамках выполняемой процедуры, является шаг соответствующего уровня, <u>иначе говоря</u>, шаг – это неделимый (законченный) набор элементарных действий.</p>
14.	Контраст (Contrast)	Однако однако		Delete Save Delete_msave	<p>Данные логики лишены недостатков с точки зрения однозначности формулируемых на их базе свойств. <u>Однако</u>, как показывает практика, их мощность позволяет формулировать лишь относительно небольшое количество однотипных условий, а этого, в свою очередь, может быть недостаточно для проверки тех или иных свойств модели конкретной системы.</p> <p>Основная проблема, возникающая в связи с этим, – гетерогенность онтологий разных источников, которая может препятствовать связыванию данных [8]. Однако существует множество исследований, с разным успехом преодолевающих эту проблему.</p>
15.		Несмотря на то, что		Delete_save	<p>Исторически типичным в таких случаях является решение <i>специальным ПО</i> (СПО) функциональных задач ИС как во взаимодействии с локальной БД, так и при вводимых оператором внешних данных и обмене специализированными сообщениями между</p>

					<p>СПО, размещенным в различных узлах ИС.</p> <p><u>Несмотря на то, что</u> значительная часть задач, решаемых в различных узлах ИС, идентична или подобна, обмен сообщениями между СПО существенно препятствует его унификации.</p> <p>В этом случае для любого человека независимо от места его проживания откроется возможность получить образование мирового класса. <i>Несмотря на то что сейчас у массовых курсов очень высокие показатели незавершенного обучения (нередко достигают даже 95 %), они обладают огромным потенциалом, требуется только более мотивирующая персонализированная поддержка [2].</i></p> <p>А такие средства, как Google Goggles или Word Lens, позволяют пользователю читать надписи на иностранном языке, просто поднеся к ним камеру телефона, на котором установлено приложение [7].</p> <p><i>Несмотря на то что очки Google Glass, на которые до релиза возлагались большие надежды, пока не позволяют пользователям получить полноценную дополненную реальность, они все же содержат дюжину датчиков, необходимых для ее реализации.</i></p>
16.		Обратите внимание		Save_delete	<p>Все этапы формирования онтологии вместе с ее оценкой можно свести к схеме, представленной на рисунке 2 [2]. <i>Обратите внимание на цикличность алгоритма: исходная, возможно, пустая онтология дополняется новыми объектами, концептами и отношениями, оценивается и затем уже используется как база для дальнейшего расширения.</i></p>
17.	Цель (Purpose)	Для того, чтобы		Save	<p>Однако необходимы еще более глубокое осмысление получаемых результатов и дополнительные исследования <u>для того, чтобы</u> с помощью</p>

					программного комплекса получать и анализировать действительно наиболее важную информацию. <u>Для того, чтобы</u> использовать любые алгоритмы машинного обучения и инструментов визуализации, они должны быть включены в первую очередь.
18.		Для того, чтобы Чтобы		Save	<u>Чтобы</u> избежать переобучения, количество обучающих примеров должно быть соразмерно числу используемых терминов.
19.		целью которого		Save	Корнелльский университет реализовал проект «Matlab on the Teragrid» [1], <u>целью которого</u> являлось предоставление Matlab пользователям Teragrid в качестве сервиса, в том числе с использованием порталов научного взаимодействия, таких как panohub.org [2].
20.		Целью данной работы		Save	<u>Целью данной работы</u> является разработка web-сервиса, автоматизирующего реализацию баз знаний продукционного типа на основе результатов концептуального (когнитивного) моделирования.
21.	Обоснование (Evidence)	Очевидно, что		Mdelete_Save	<u>Очевидно, что</u> критическим аспектом приведенной классификации является соотношение операционной нагрузки и локальных вычислительных возможностей.
22.		Таким образом,		Mdelete_Save	<u>Таким образом,</u> все сервисные операции с ЭБД выполняются автоматически, без участия экипажа.
23.		можно сделать вывод			<i>Рассмотрев различные варианты практических задач по оптимальному расположению грузов и выделив сходства и различия между ними и задачей оптимального размещения грузов на борту транспортного грузового корабля, <u>можно сделать вывод</u>, что универсального метода решения задачи</i>

					оптимального размещения не существует, в каждой конкретной задаче есть свои особенности и ограничения, которые необходимо учитывать.
24.		Можно сделать вывод		Save	<u>Можно сделать вывод</u> о сложности данной темы и необходимости усовершенствовать преподнесение материала в рамках семинарских занятий.
25.		Как показала практика Как показывает практика Как показали эксперименты ...		Save	<u>Как показала практика</u> , таким инструментом может быть простая таблица, содержащая два столбца: в одном указываются задачи ТЗ, в другом – соответствующие им прецеденты (табл. 1).
26.		Важной составляющей		Save	<u>Важной составляющей</u> имитационно-тренажерных комплексов является система управления.
27.		Преимуществами системы		Save	<u>Преимуществами системы</u> являются простота ее использования, нетребовательность к ресурсам и расширяемость.
28.		В трактовке		Mdelete_Save	<u>В трактовке</u> стандарта POSIX-2001 в трассировке логически участвуют три процесса, которые физически могут совпадать между собой: трассируемый (целевой), трассирующий (управляющий трассировкой) и анализирующий данные трассировки.
29.		Исследования показали		Save	<u>Исследования показали</u> , что наилучший результат получается при удалении всей иерархии внутри блока перед синтезом.
30.		Основной идеей		Save	<u>Основной идеей</u> технологии кеинга является выделение объекта от однородного фона.
31.		В настоящей работе		Save	<u>В настоящей работе</u> развивается метод математического и компьютерного моделирования интервально стохастических температурных полей, обусловленных интервальным стохастическим

					характером входных данных, определяющих тепловые режимы технической системы.
32.		В данной работе		Save	<u>В данной работе</u> используется формальный язык для описания тестовых данных «Sulley», специально разработанный для тестирования приложений рабочей группой Университета Тулейна (США) и позволяющий описывать процедуру анализа с необходимым уровнем детализации [9].
33.		В данной статье		Save	<u>В данной статье</u> описывается реализация генетического алгоритма для выявления и отбора наиболее релевантных результатов, полученных в ходе последовательно выполняемых операций тематического поиска.
34.		На практике приведен		Save	На практике приведенный алгоритм необходимо модифицировать прокладкой перекрестных маршрутов между всеми <i>процессорными элементами</i> (ПЭ) и ограничениями на просмотр портов коммутаторов (отдельные крайты в сложной системе могут включаться неодновременно) – необходима локализация алгоритма в крайте или в группе крайтов. Кроме того, совершенно не учитываются предполагаемые потоки данных между ПЭ.
35.		Прежде всего			Прежде всего применяются учебно-прикладные игры, воспроизводящие трудовые процессы специалистов ракетно-космической отрасли (космонавтов, работников центра управления полетами и т.п.), а также игры, развивающие интеллектуальные способности.

36.	Причина (Cause-Effect)	Так как		Save_delete	Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ, <i>так как в некоторых ситуациях эпизоды могут быть настолько продолжительными, что задержки процесса обучения, связанные с необходимостью завершения эпизодов, будут слишком велики.</i>
37.		Поэтому		Mdelete_save	<u>Поэтому</u> техническая система, созданная из различных серийно изготавливаемых элементов, также будет иметь параметры и характеристики, носящие неопределенный характер и изменяющиеся в пределах некоторых интервалов.
38.	Фон (Background)	При этом		Save_delete	В силу интервально стохастического характера параметров и характеристик технической системы решение уравнений стохастической математической модели, описывающей температурное поле, будет интервально стохастическим полем $T(\omega) = T(x, y, z, \omega)$. <u>При этом</u> температура в каждой точке технической системы будет изменяться в некотором интервале и иметь распределение вероятностей, вообще говоря, отличное от равномерного.
39.		Для его внедрения /использования		save	Для его внедрения в единое синтезированное трехмерное окружение создан метод рир-проекции, базирующийся на методе 3D-кеинга.

Приложение Б. Шаблоны для сглаживания

Шаблоны для дополнения

«Введение»

$X \in \{\text{В статье, В работе, ...}\}$

$Y_V \in \{\text{рассматриваются, рассматривается, ...}\}$

$Y_N \in \{\text{задачи, метод, способ, подходы, ...}\}$

Z – оставшаяся часть предложения (сателлит).

«Новизна»

$X \in \{\text{Новизна, Новизна и перспективность, ...}\}$

$Y_N \in \{\text{метода, алгоритма, подходов, ...}\}$

$Y_V \in \{\text{заключается, определяется, ...}\}$

Z – оставшаяся часть предложения (сателлит).

«Цель»

Вариант 1

$X_p - \{\text{Целью, Основной целью, Основным направлением, ...}\}$

$X_w - \{\text{данной работы, статьи, исследования, модели, ...}\}$

$Y_V - \{\text{является, играет, занимает, считается, ...}\}$

KW – ключевые слова;

Z – оставшаяся часть предложения (сателлит с маркером или без маркера).

Вариант 2

$Y_V \in \{\text{Показана, Представлена, Исследуется, ...}\}$

$Y_N \in \{\text{целесообразность взаимодействия}\}$

KW – ключевые слова

$T \in \{\text{с системой, на основе, по вопросам, ...}\}$

Z – оставшаяся часть предложения (сателлит с маркером или без маркера).

Вариант 3

$Y_N \in \{\text{Применение, Использование, Разработка, Вычисление, ...}\}$

$X_w \in \{\text{данной работы, статьи, исследования, модели, этого, ...}\}$

$K_p \in \{\text{полезно для}\}$

KW – ключевые слова

$P \in \{\text{с целью, ...}\}$

$P_p \in \{\text{формирования, обеспечения, улучшения, верификации модели, ...}\}$

Z – оставшаяся часть предложения (сателлит с маркером или без маркера).

«Метод» (Методика | Планирование | Методология | Модель | Стратегия | Подход | Оценка | Определение | Формирование | Анализ | Проектирование)

Вариант 1

$Y_V \in \{\text{Рассматриваются, Проводятся, Перечислены, Предлагаются, ...}\}$

$Y_N \in \{\text{методы, методика, Система, возможности, задачи, ...}\}$

$O \in \{\text{где основой являются, где используются, ...}\}$

Z – оставшаяся часть предложения (сателлит с ключевыми словами или без них).

Вариант 2

$X \in \{\text{В статье, В данной работе, В данной статье, В модели, В информационных системах, ...}\}$

$Y_V \in \{\text{Рассматриваются, Проводятся, Перечислены, Предлагаются, ...}\}$

KW – ключевые слова

$T \in \{\text{где применяются, с применением, ...}\}$

KW – ключевые слова

$K_d \in \{\text{каждое из которых, которые, примером являются, с применением, ...}\}$

KW – ключевые слова

Вариант 3

$X \in \{\text{Создание, Применение, Использование, Разработка, Вычисление, ...}\}$

$Y_N \in \{\text{метода, методики, системы, ...}\}$

$O \in \{\text{где основой являются, где используются}\}$

Z – оставшаяся часть предложения (сателлит с ключевыми словами или без них).

«Реализация»

Вариант 1

$Y_N \in \{\text{Алгоритм, системы, ...}\}$

$Y_V \in \{\text{реализован, реализованы, ...}\}$

$PREP \in \{\text{на, в, ...}\}$

$KW \in \{\text{языке C++, ...}\}$

Вариант 2

$Y_V \in \{\text{Описана, ...}\}$

$Y_N \in \{\text{Программная реализация, программное обеспечение, ...}\}$

$K_r \in \{\text{разработанного алгоритма, ...}\}$

Z – оставшаяся часть предложения (сателлит с маркером или без маркера).

«Недостатки»
 $Y_N \in \{\text{Недостаток, Достоинства, ...}\}$
 $N \in \{\text{методов, ...}\}$
 $PREP \in \{\text{в том, что; ...}\}$
 $Y_V \in \{\text{рассматривают, ...}\}$
 Z – оставшаяся часть предложения (спутник с маркером или без маркера).
«Заключение»
 $Y_V \in \{\text{Приведены, Рассмотрены, ...}\}$
 KW – ключевые слова

 $K_C \in \{\text{таким образом, чтобы; где можно сделать вывод, ...}\}$
 Z – оставшаяся часть предложения (спутник).

Таблица 20 – Примеры использования шаблонов для дополнения

Тип шаблона «Введение»				
В настоящей работе	описываются	вопросы проектирования функциональных проблем,	основанные на аппарате искусственного интеллекта.	
В рамках проблемы	используют	только два подхода	к формированию оценок данного алгоритма,	показывающих только сильные и слабые стороны.
Тип шаблона «Цель»				
Целью исследования	является	оптимизация распределения ресурсов	среди уязвимых	с точки зрения временных задержек и скорости обработки запросов внешних пользователей.
С целью	учета опыта, навыков, компетенций и предпочтений сотрудников организации в системах документооборота	предлагается использовать	базы знаний компетенций специалистов,	которые можно формализовать с помощью онтологий.
Тип шаблона «Метод»				

Рассматривается	алгоритм ImpAA,	осуществляющий поиск минимальных абдуктивных объяснений		с помощью первичных импликат.	
Методы	можно применять	к пространствам		с различной дискретной математической структурой.	
Предложена	методика,	позволяющая определять показатели качества обнаружения РЛС	для широкого класса моделей сигналов,	в качестве малозаметных и малоразмерных целей в условиях стационарных гауссовских, шумовых импульсных помех,	а также в беспомеховой обстановке.
Тип шаблона «Недостатки»					
Основные достоинства	парадигмы квантовых дискретных информационных динамических систем	как информационных систем и новых форм компьютинга в технологиях компьютерного и математического моделирования		закключаются	в том, что они с позиций единой концептуальной схемы позволяют естественным образом учитывать следующее.
Основной недостаток	метода радиационного контроля в том, что рассеянное излучение	в зависимости от		энергии первичного излучения измеряет качество снимка, снижает контрастность и четкость изображения,	а следовательно, и чувствительность самого метода.
Тип шаблона «Реализация»					
Система	реализована	на платформе .NET Framework 2.0.			
Реализованы	механизмы	подготовки шаблонов моделей		и генерации кода конечной модели.	
Данный лабораторный практикум	представляет собой реализацию	в системе Matlab совокупности численных методов		для вычисления функций с заданной точностью,	для решения нелинейных уравнений, системы линейных алгебраических уравнений и дифференциальных уравнений.
Тип шаблона «Новизна»					
Эффективность	комбинированного метода	подтверждается экспериментами		на текстовой коллекции отзывов о фильмах	семинара РОМИП-2011.
Подчеркивается	научная новизна	ожидаемых результатов,		а также определена целевая аудитория конечных пользователей инструментальных программных средств.	

Тип шаблона «Заключение»			
Полученные	результаты	позволяют	говорить об эффективности предложенной методики.
Приведены	результаты	расчетов размеров и формы вихревых следов	для различных типов воздушных судов.

Таблица 21 – Шаблоны для удаления

№	До сглаживания	После сглаживания
1.	Рассматриваем	Рассмотрено
2.	Можно сделать вывод	Сделан вывод
3.	На рисунке 4 представлен	Представлен
4.	Для этого предлагается	Предложено
5.	Рассмотрим	Рассмотрено рассмотрен рассмотрены рассмотрена
6.	Были апробированы	Апробировано
7.	В статье приведен	приведено / привели / Преведен
8.	На основе данных рассуждений создан	Создан
9.	В нашей статье	В статье
10.	В трактовке	Дается краткий обзор
11.	Честно говоря	Вообще
12.	Важной составляющей	Составляющей
13.	Основной идеей	Основой
14.	Как показала практика	-
15.	Как показывает практика	-
16.	Как показали эксперименты	-
17.	Как показывает опыт	-
18.	Целью данной работы	Целью работы
19.	С точки зрения схемотехнического проектирования важную роль	Важную роль
20.	В статье рассмотрены	Рассмотрены
21.	В дальнейших исследованиях и разработках планируется рассмотреть	Планируется рассмотреть
22.	Данная операция	операция
23.	Подытоживая сказанное,	-

Приложение В. Примеры работы системы

1. Список тем, найденных в документе

Название документа:

«Об одном подходе к оценке качества обработки видеографической информации»

topic_0 | 0,13593 [Эталонный, изображение, метод, граница, качество, работа, программа, реализация, обработка, задача, оценка, результат, выделение, программный, объект, информация, рассматривать, видеографический, являться, набор, решение, точка, пиксел, контур, система, получать, область, мера, функция, метрика, сегментатор, шум, деградация, ground truth, универсальный, использовать, величина, эталонное изображение, функция принадлежности, метод обработки видеографической информации, качество работы программ, универсальная оценка качества, программная реализация саппу, аффинные преобразования]

topic_1 | 0,08513 [Текстурный, получение, граничные точки, фон, принадлежность, выбор, ситуация, саппу, позволять, основа, подход, показывать, плотность, исследование, расстояние, истинный, контролировать, особенность, искусственный, эталон, отдел, реставрация, уточнение, сложный, разрабатывать, квадрат, конкретный, ось, зашумление, программные реализации, плотность локальных экстремумов]

topic_2 | 0,03442 [Специфичность, выявлять, называть, описывать, контраст, отмечать, определять, изменяться, отличие, линия, экстремум, равный, решающий, сегментация, хаусдорф, локальный, средство, относительно, maxdif, класс, идеология, формирование, угловой, левый, создавать, зависимость, понимание, связь, углубление, работа программ, оценка качества, измерение качества, выбор программ, реставрация изображений]

topic_3 | 0,065789 [Поведение, jseg, образ, известный, smith, вариация, обладать, материал, шах, высокий, помощь, кривизна, содержать, изменение, возможность, технический, распознавание, искажение, рамка, реализовать, прикладной, абсцисса, petra, гауссов, скачок, откладывать, эталонный, неформальный, правый, бестекстурный, подвергать, gothwell, левая часть рисунка, ось абсцисс, мера отличия, база эталонных изображений, статистическая обработка результатов]

2. Промежуточный результат риторического анализа

- 1: _ В данной РАБОТЕ описывается ПОДХОД к обработке видеографической ИНФОРМАЦИИ, сложившийся к настоящему времени в ОТДЕЛЕ ОБРАБОТКИ и РАСПОЗНАВАНИЯ видеографической ИНФОРМАЦИИ НИИСИ РАН.
- 3: _ На ОСНОВЕ созданной в ОТДЕЛЕ 3D-модели отображения земной поверхности в реальном масштабе времени [1] был разработан многомашинный макет автоматизированной СИСТЕМЫ мониторинга земной поверхности ДЕДАЛ [2], предназначенной для дистанционного обнаружения и РАСПОЗНАВАНИЯ движущихся ОБЪЕКТОВ.
- 4: _ Необходимо также отметить разработанную компьютерную систему ПРИЗМА [3], позволяющую по заданному НАБОРУ изображений и эталонов подбирать МЕТОДЫ их ОБРАБОТКИ,.
- 8: _ Это требование может быть удовлетворено, если все МЕТОДЫ оцениваются на одном и том же видеографическом МАТЕРИАЛЕ.
- 14: _ Эталонные ИЗОБРАЖЕНИЯ должны содержать максимально полный НАБОР элементов ИЗОБРАЖЕНИЯ, являющихся типовыми для ЗАДАЧИ, решаемой рассматриваемыми МЕТОДАМИ ОБРАБОТКИ видеографической ИНФОРМАЦИИ.
- 23,24: _ Для эталонных изображений, подобранных в соответствии с описанными ПРИНЦИПАМИ, в КАЧЕСТВЕ универсальной ОЦЕНКИ качества решения ЗАДАЧИ ОБРАБОТКИ видеографической ИНФОРМАЦИИ можно взять некоторую меру отличия РЕЗУЛЬТАТОВ ОБРАБОТКИ этой ИНФОРМАЦИИ от ground truth. Необходимо отметить, что ВЫБОР конкретной меры определяет содержательную интерпретацию получаемых ОЦЕНОК. . В частности, можно брать меры отличия, полученные на основе метрик ЕВКЛИДА, ХАУСДОРФА, статистических, нечетких мер и т.п.
- 33: _ Вместе с тем описанные СИТУАЦИИ являются вполне типичными для естественных изображений.
- 51: _ Обработанное изображение ближе к ground truth, чем заШУМленное.

- 55:_ Эта задача обычно решается с ПОМОЩЬЮ ПРОГРАММ на ОСНОВЕ так называемого МЕТОДА активного контура [6], для РЕАЛИЗАЦИЙ которого трудными являются СИТУАЦИИ, когда контур объекта имеет большую кривизну.
- 56:_ Поэтому в НАБОР эталонных изображений для ОЦЕНКИ качества РАБОТЫ соответствующих ПРОГРАММ уточнения контуров были включены контуры с широким диапазоном изменений кривизны.
- 58:_ Следует отметить, что типичной ситуацией, влияющей на РЕЗУЛЬТАТЫ РАБОТЫ ПРОГРАММ, решающих задачу уточнения контуров ОБЪЕКТОВ, является СЛОЖНОСТЬ ФОНА.
- 59:_ Поэтому к НАБОРУ эталонных контуров добавляются и образцы ФОНА.
- 66:_ В КАЧЕСТВЕ эталонных изображений естественно было взять ИЗОБРАЖЕНИЯ, использованные при ИССЛЕДОВАНИИ ПРОГРАММ выделения границ, а аналогом ДЕГРАДАЦИИ в рассматриваемом СЛУЧАЕ являются собственно аффинные преобразования.
- 75:_ Одним из сложных СЛУЧАЕВ для СЕГМЕНТАТОРОВ является наличие УГЛОВ на ИЗОБРАЖЕНИИ.
- 76:_ И такие СИТУАЦИИ нельзя считать исключительными.
- 82:_ Как видим, чем острее УГОЛ, тем больше могут быть ИСКАЖЕНИЯ.
- 91,92:_ Как видим, только ПРОГРАММная реализация Canny позволила выявить все УГЛОВые точки, являющиеся узловыми для данного ИЗОБРАЖЕНИЯ. Однако при использовании ОЦЕНОК, и качество РЕЗУЛЬТАТОВ РАБОТЫ неразлично.
- 93:_ Использование классических метрик не выявляет преимущество ПРОГРАММной реализации МЕТОДА Canny, не пропустившей УГЛОВые точки квадрата.
- 97:_ Если эти функции принадлежности будут подчеркивать значимость пикселей в особенностях ГРАНИЦЫ объекта, то нечеткие МЕТРИКИ должны уловить различие в РАБОТЕ ПРОГРАММ, выделяющих ГРАНИЦЫ, относительно этих особенностей.
- 102:_ Отметим, что в рассматриваемом ПРИМЕРЕ функция принадлежности РЕЗУЛЬТАТОВ РАБОТЫ ПРОГРАММ является вырожденной, принимающей значение 1 только на определенных ПРОГРАММой граничных ПИКСЕЛАХ.
- 105:_ Можно утверждать, что использование нечетких мер сходства и расширение понятия эталонных изображений до нечетких позволяют более полно выявлять ОСОБЕННОСТИ сравниваемых ПРОГРАММ.
- 106:_ Рассмотрим применение идеологии получения универсальной ОЦЕНКИ качества РАБОТЫ различных ПРОГРАММных РЕАЛИЗАЦИЙ методов, используемых при решении задач текстурного анализа,, наПРИМЕР, задачу выделения на ИЗОБРАЖЕНИИ текстур.
- 108:_ На РИСУНКЕ 15 приведен ПРИМЕР из НАБОРА искусственных эталонных изображений.
- 109:_ , чтобы в пределах текстурных областей могли меняться размер текстуры, а также контраст ГРАНИЦЫ между текстурной и бестектурной ОБЛАСТЯМИ.

3. Квазиреферат

- 1: В данной работе описывается подход к обработке видеографической информации, сложившийся к настоящему времени в отделе обработки и распознавания видеографической информации НИИСИ РАН.
Weight = 0.088
описывать: 4
подход: 5
- 3: На основе созданной в отделе 3D-модели отображения земной поверхности в реальном масштабе времени [1] был разработан многомашинный макет автоматизированной системы мониторинга земной поверхности ДЕДАЛ [2], предназначенной для дистанционного обнаружения и распознавания движущихся объектов.
Weight = 0.132
создавать: 4
разрабатывать: 4
На основе: 5
- 4: Необходимо также отметить разработанную компьютерную систему ПРИЗМА [3], позволяющую по заданному набору изображений и эталонов подбирать методы их обработки,.
Weight = 0.120
позволять: 3
отмечать: 4
разрабатывать: 4
- 8: Это требование может быть удовлетворено, если все методы оцениваются на одном и том же видеографическом материале.
Weight = 0.022
оценивать: 2

14: Эталонные изображения должны содержать максимально полный набор элементов изображения, являющихся типовыми для задачи, решаемой рассматриваемыми методами обработки видеографической информации.

Weight = 0.196

рассматривать: 5

содержать: 4

являть: 5

решать: 4

23,24: Для эталонных изображений, подобранных в соответствии с описанными принципами, в качестве универсальной оценки качества решения задачи обработки видеографической информации можно взять некоторую меру отличия результатов обработки этой информации от ground truth. Необходимо отметить, что выбор конкретной меры определяет содержательную интерпретацию получаемых оценок. В частности, можно брать меры отличия, полученные на основе метрик Евклида, Хаусдорфа, статистических, нечетких мер и т.п.

Weight = 0.120

отмечать: 4

описывать: 4

определять: 3

33: Вместе с тем описанные ситуации являются вполне типичными для естественных изображений.

Weight = 0.098

описывать: 4

являть: 5

55: Эта задача обычно решается с помощью программ на основе так называемого метода активного контура [6], для реализаций которого трудными являются ситуации, когда контур объекта имеет большую кривизну.

Weight = 0.200

называть: 2

решать: 4

задача: 4

являть: 5

на основе: 5

58: Следует отметить, что типичной ситуацией, влияющей на результаты работы программ, решающих задачу уточнения контуров объектов, является сложность фона.

Weight = 0.098

отмечать: 4

являть: 5

66: В качестве эталонных изображений естественно было взять изображения, использованные при исследовании программ выделения границ, а аналогом деградации в рассматриваемом случае являются собственно аффинные преобразования.

Weight = 0.152

использовать: 4

рассматривать: 5

являть: 5

75: Одним из сложных случаев для сегментаторов является наличие углов на изображении.

Weight = 0.054

являть: 5

91,92: Как видим, только программная реализация Canny позволила выявить все угловые точки, являющиеся узловыми для данного изображения. Однако при использовании оценок, и качество результатов работы неразлично.

Weight = 0.132

позволять: 3

реализация: 5

являть: 5

93: Использование классических метрик не выявляет преимущество программной реализации метода Canny, не пропустившей угловые точки квадрата.

Weight = 0.045

Использование: 5

102: Отметим, что в рассматриваемом примере функция принадлежности результатов работы программ является вырожденной, принимающей значение 1 только на определенных программой граничных пикселах.

Weight = 0.152

отмечать: 4

являть: 5

- рассматривать: 5
 105: Можно утверждать, что использование нечетких мер сходства и расширение понятия эталонных изображений до нечетких позволяют более полно выявлять особенности сравниваемых программ.
 Weight = 0.077
 позволять: 3
 использование: 5
 106: Рассмотрим применение идеологии получения универсальной оценки качества работы различных программных реализаций методов, используемых при решении задач текстурного анализа, например, задачу выделения на изображении текстур.
 Weight = 0.142
 применение: 5
 использовать: 4
 рассматривать: 5

4. Сравнение авторской и автоматически полученной аннотации

Таблица 22 – Сравнение автоматически полученной и авторской аннотаций

Жирным шрифтом выделены совпадающие фрагменты аннотаций.

Авторская аннотация	Автоматическая аннотация
<p>В статье изложена разработанная в НИИСИ РАН идеология построения универсальной оценки качества работы компьютерных программ, реализующих тот или иной метод решения некоторой задачи обработки видеографической информации. Такая оценка позволяет сравнивать на сопоставимой основе эффективность работы программ с целью выбора среди них наиболее адекватных условиям применения. Разработчикам практических систем обработки видеографической информации изложенный подход позволит уже на стадии проектирования системы сделать обоснованный выбор программной реализации для решения стоящей перед ними задачи. Рассмотрены примеры применения изложенного подхода для сравнительной оценки ряда широко используемых программных реализаций решения задач выделения границ, реставрации, уточнения контуров, сегментации, текстурного анализа, а также исследование их устойчивости относительно аффинных преобразований. Внешние условия моделировались зашумлением и размытием стандартизованного набора эталонных изображений. В роли универсальной оценки качества в примерах были рассмотрены статистические и размытые меры, метрики Евклида и Хаусдорфа. Эти примеры позволили выявить особенности поведения программных реализаций и получить области их предпочтительного применения.</p>	<p>В данной работе описывается подход к обработке видеографической информации, сложившийся в отделе обработки и распознавания видеографической информации НИИСИ РАН. На основе созданной в отделе 3D-модели был разработан многомашинный макет, предназначенной для дистанционного обнаружения и распознавания движущихся объектов, реализующий тот или иной метод решения некоторой задачи обработки видеографической информации. Эталонные изображения должны содержать максимально полный набор элементов изображения, являющихся типовыми для задачи, решаемой рассматриваемыми методами обработки видеографической информации. Необходимо отметить, что выбор конкретной меры определяет содержательную интерпретацию получаемых оценок. Изложенный подход решается с помощью программ на основе так называемого метода активного контура [6], для реализаций которого трудными являются ситуации, когда контур объекта имеет большую кривизну. В качестве эталонных изображений естественно было взять изображения, использованные при исследовании программ выделения границ, а аналогом деградации в рассматриваемом случае являются собственно аффинные преобразования. Необходимо отметить, что выбор конкретной меры определяет содержательную интерпретацию получаемых оценок. В частности, можно брать меры отличия, полученные на основе метрик Евклида, Хаусдорфа, статистических, нечетких мер и т.п.</p>

Приложение Г. Свидетельства о регистрации программ ЭВМ

Blank area for the certificate content.

РОССИЙСКАЯ ФЕДЕРАЦИЯ



СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2018614456

**Морфологическая система «Стемматизация и генерация
словоформ казахского языка»**

Правообладатель: *Федеральное государственное бюджетное
учреждение науки Институт вычислительных технологий
Сибирского отделения Российской академии наук (ИВТ СО РАН)
(RU)*

Автор: *Бакиева Айгерим (KZ)*



Заявка № **2017663195**

Дата поступления **19 декабря 2017 г.**

Дата государственной регистрации
в Реестре программ для ЭВМ **06 апреля 2018 г.**

*Руководитель Федеральной службы
по интеллектуальной собственности*

Г.П. Илизаров **Г.П. Илизаров**

СВИДЕТЕЛЬСТВО

о государственной регистрации прав
на объект авторского права

№

443

17 апреля 2017

г.

Настоящим удостоверяется, что в Министерстве юстиции Республики Казахстан зарегистрированы исключительные имущественные права на объект авторского права под названием «Қазақ және түрік тілдеріне арналған LINK GRAMMAR PARSER синтаксистік талдағышы» (программа для ЭВМ), авторами которого по заявлению авторов являются Еримбетова Айгерим Сембековна, Батура Татьяна Викторовна, Мурзин Фёдор Александрович, Сагнаева Сауле Кайроллиевна, Бакиева Айгерим Муратовна.

По заявлению авторов исключительные имущественные права на объект авторского права, созданный в период с 1 ноября 2014 года по 16 октября 2016 года, принадлежат Еримбетовой А.С., Батура Т.В., Мурзину Ф.А., Сагнаевой С.К., Бакиевой А.М. и авторы гарантируют, что при создании вышеуказанного объекта не были нарушены права интеллектуальной собственности других лиц.

Запись в реестре за № 443 от 17 апреля 2017 года имеется.

Заместитель министра

Э. Азимова

СВИДЕТЕЛЬСТВО

ИС 008065

Приложение Д. Акты о внедрении

ФАНО РОССИИ



ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ
 СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК
 (ИВТ СО РАН)

АКТ

о внедрении результатов диссертационного исследования

06.03.2017

№ 1

Новосибирск

Настоящий акт подтверждает, что научные и практические результаты, полученные в ходе подготовки диссертации на соискание ученой степени кандидата технических наук по специальности 05.13.17 «Теоретические основы информатики» аспирантки факультета информационных технологий Новосибирского государственного университета А.М.Бакиевой «Стемматизация и генерация словоформ казахского языка для систем автоматической обработки текстов», внедрены в Институте вычислительных технологий СО РАН. Предложенные А.М.Бакиевой модели и алгоритмы стемматизации и генерации словоформ казахского языка были реализованы ею в форме веб-приложения, опубликованного Институтом вычислительных технологий СО РАН в виде научного ИТ-сервиса, доступного широкому кругу пользователей по адресу http://poem.ict.nsc.ru/~bakieva_aigerim/kazGen/. Сервис автоматизирует обработку текстов (в частности, поэтических) на казахском языке и представляет интерес для специалистов в области филологии, занимающихся изучением казахского языка и анализом текстов на казахском языке.

Врио директора Института
 к.ф.-м.н.



А.В.Юрченко

РОССИЯ
Новосибирск
Общество с ограниченной ответственностью
“Новые программные системы”

ИНН 5408254808, ОГРН 1075473011900

630090, г.Новосибирск, пр. Лаврентьева, д.6, оф.222

УТВЕРЖДАЮ
Директор ООО

«Новые программные системы»

Д.Н. Штокало

18.12.2018 г.



АКТ

о внедрении научно-исследовательских результатов диссертационной работы
Бакиевой Айгерим Муратовны по теме «Модели определения тем текстов,
основанные на графах, и их применение для решения задачи
автореферирования»

Настоящий акт подтверждает, что результаты диссертационного исследования по теме «Модели определения тем текстов, основанные на графах, и их применение для решения задачи автореферирования», полученные соискателем Бакиевой Айгерим Муратовны по специальности 05.13.17 – «Теоретические основы информатики» применяются в ООО «Новые программные системы» в процессе проведения научных исследований для анализа текстовой информации.

Бакиевой А.М. разработаны и реализованы следующие методы и алгоритмы: метод построения аннотаций научных текстов, использующий представление текстов в виде графов; алгоритм обнаружения важных элементов текста на основе теории риторических структур; алгоритм построения расширенных тематических моделей и выделения многословных терминов. Программное обеспечение может применяться для работы с текстами на русском, казахском и турецком языках.

Бакиевой А.М. реализован обширный набор программных инструментов, предназначенный для поддержки проводимых исследований и представляющий практический интерес.

Члены комиссии

к.ф.-м.н.

к.б.н.

Д.С. Мигинский

Д.В. Антоненц

«УТВЕРЖДАЮ»

директор Института систем информатики
им. А.П. Ершова СО РАН

к.ф.-м.н. _____ А.Ю. Пальянов

«21» _____ 2018 г.

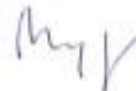
А К То внедрении научно-исследовательских результатов диссертационной работы
Бакиевой Айгерим Муратовны

Настоящий акт подтверждает, что научные и практические результаты, полученные соискателем Бакиевой Айгерим Муратовной в ходе выполнения диссертационной работы по теме «Модели определения тем текстов, основанные на графах, и их применение для решения задачи автореферирования» по специальности 05.13.17 – Теоретические основы информатики, используются в Лаборатории моделирования сложных систем федерального государственного бюджетного учреждения науки Института систем информатики им. А.П. Ершова СО РАН.

1. Созданный программный комплекс применяется для анализа больших наборов данных с целью автоматического извлечения важной информации по перспективным научным направлениям и технологиям. Данные представляют собой наборы до 60 тысяч файлов.

2. Предложенные А.М. Бакиевой модели тем текстов (униграммные и расширенные), алгоритм обнаружения наиболее содержательных элементов текста на основе теории риторических структур, алгоритм построения аннотаций и лингвистическая база знаний используются как встраиваемые компоненты при реализации различных проектов.

Заместитель директора
по научной работе ИСИ СО РАН,
зав. лаб. моделирования сложных систем
к.ф.-м.н.



Ф.А. Мурзин