Т. В. Батура, А. М. Бакиева

МЕТОДЫ И СИСТЕМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Новосибирск 2019

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ИНСТИТУТ СИСТЕМ ИНФОРМАТИКИ ИМ. А. П. ЕРШОВА СО РАН

Т. В. Батура, А. М. Бакиева

МЕТОДЫ И СИСТЕМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Монография

Новосибирск
2019

УДК 004.8+519.76+519.688
ББК 32.81:32.97
Б 287

Рецензент
д-р техн. наук, доц. *В. Б. Барахнин*

Батура, Т. В.
Б 287 Методы и системы автоматического реферирования текстов : монография / Т. В. Батура, А. М. Бакиева ; Ин-т систем информатики им. А. П. Ершова СО РАН. — Новосибирск : ИПЦ НГУ, 2019. — 110 с.

ISBN 978-5-4437-0974-1

В монографии рассмотрены различные принципы построения алгоритмов и систем автореферирования, позволяющие извлекать ценную информацию из текстовых документов. Описан новый метод автоматического составления рефератов научно-технических текстов, основанный на риторическом анализе и тематическом моделировании. Предложенный метод сочетает в себе использование лингвистической базы знаний, графового представления текстов и машинного обучения.

Монография предназначена для широкого круга читателей с любым уровнем подготовки, интересующихся теоретическими исследованиями и практическими приложениями, связанными с обработкой и интеллектуальным анализом текстовой информации.

The monograph discusses various principles of constructing algorithms and autosummarization systems that allow to extract valuable information from text documents. A new method for the summarization of scientific and technical texts based on rhetorical analysis and topic modeling is described. The proposed method combines the use of a linguistic knowledge base, graph representation of texts, and machine learning.

The monograph is intended for a wide range of readers with any level of knowledge, who are interested in theoretical research and practical applications related to the natural language processing and intellectual analysis of text information.

УДК 004.8+519.76+519.688
ББК 32.81:32.97

Исследования выполнены при финансовой поддержке РФФИ
в рамках научного проекта № 19-07-01134.

ISBN 978-5-4437-0974-1

ВВЕДЕНИЕ

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 попрежнему 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 их тем.

Нашей главной целью является создание новых методов и описание формальных моделей, применяемых для решения задачи автореферирования. В данной монографии предложен новый 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 становится особенно актуальным, поскольку информационные технологии используются практически во всех отраслях науки и техники.

В настоящее время наблюдается растущий научный интерес к области автоматизации реферирования и аннотирования. Этой пробле-

мой начали заниматься во второй половине XX века такие ученые, как Г. П. Лун, Д. Марку, К. Оно, У. Хан, Д. Радев, Г. Саггион, Л. Плаза, Г. П. Эдмундсон, Дж. Купитс, Е. Льорет, Дж. Поллок, Т. Стрлзалковски. 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 и 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.

На основе предложенного гибридного метода и описанных моделей создана система автоматического реферирования документов. Разработанное программное обеспечение может применяться для построения систем машинного понимания текста, систем автоматической обработки текстов, информационно-поисковых систем и других информационных систем, основанных на знаниях. Отметим, что в данный момент исследования ориентированы на русскоязычные источники. Однако предложенный метод планируется адаптировать также для работы с тюркскими языками, такими как казахский и турецкий.

В главе 1 приведен обзор существующих методов автоматического реферирования, перечислены отечественные и зарубежные программные продукты, реализующие некоторые из методов. Рассмотрены

две классификации. Согласно первой выделяют три направления: экстракцию, абстракцию, гибридный подход, вторая же прибегает к разделению на пять групп методов: статистические, алгебраические, графовые, когерентные, методы на основе машинного обучения.

В главах 2, 3 и 4 подробно рассмотрены экстрагирующие, абстрагирующие и гибридные методы соответственно. Проанализированы преимущества и недостатки каждой группы методов, сделаны выводы о целесообразности их использования. Кроме того, глава 4 содержит описание предлагаемого гибридного метода автоматического построения аннотаций. 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.

Глава 5 посвящена построению униграммных моделей тем текстов и коллекций документов. Тема (для научного текста) — набор терминов, которые описывают предметную область или ее часть. Тематическая модель — набор тем в тексте. Униграммная тематическая модель — модель, в которой темы описаны однословными терминами. Расширенная тематическая модель — модель, в которой темы описаны не только однословными, но и многословными терминами. В предложенной главе проведено сравнение основных подходов тематического моделирования: *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). Для выбора наиболее подходящего метода кроме стандартных метрик оценки моделей (перплексия, доля фоновых слов, средняя контрастность ядер тем и др.) использовались такие критерии, как применимость к большим наборам данных, единственность и устойчивость решения, возможность учитывать многословные термины.

Инструмент тематического моделирования был выбран не случайно. На сегодняшний день основное требование к тематическим

моделям заключается в том, что они должны хорошо поддаваться интерпретации, чтобы конечному пользователю были 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, тематическое моделирование позволяет решать такие сложные проблемы, как синонимия и полисемия, так как учитывает контекст. Поэтому синонимы с большой долей вероятности будут отнесены к одной теме, а омонимы (слова одинаковые по написанию, но разные по значению) встретятся в разных темах.

В главе 6 рассматривается проблема многословных терминов. Перечислены трудности, возникающие при построении тематических моделей, содержащих многословные термины. Предложено решение этой проблемы с помощью набора лексико-синтаксических шаблонов.

Глава 7 содержит подробное описание алоритма построения расширенных тематических моделей. Алгоритм состоит из следующих шагов: предобработка текста; лемматизация; построение морфологического словаря при помощи *Mystem*; извлечение *n*-грамм слов с помощью алгоритма *RAKE;* согласование словосочетаний; построение тематических моделей с использованием алгоритма *ARTM*; построение расширенной тематической модели; построение словаря с весами при помощи *TF-IDF*; распределение тем и ключевых терминов по документам.

Глава 8 посвящена риторическому анализу текстов. Основными понятиями в теории риторических структур являются элементарные дискурсивные единицы (ЭДЕ) и отношения. Можно определить два типа ЭДЕ. Один из них, называемый ядром, считается наиболее важной частью высказывания, другой, сателлит, поясняет ядро и считается вторичным. Ядро содержит основную информацию, в то время как сателлит — дополнительную информацию о ядре.

Согласно теории риторических структур, любой текст может быть представлен в виде графа $G = <V, E>$, узлами V которого являются ЭДЕ, а ребрами E — риторические отношения между ними. Преоб-

разования с графами риторических структур позволяют получить квазиреферат. 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, который будет являться 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 на данном этапе, используется аппарат логики предикатов.

В главе 9 описан заключительный этап — сглаживание текста автоматически составленной аннотации. Сглаживание — процедура преобразования текста, позволяющая получить связный текст из разрозненных фрагментов и при необходимости дополнительно сократить его. Используются шаблоны двух типов: для удаления фрагментов предложений и для дополнения. При этом важно, чтобы были выполнены определенные условия для выбора подходящих шаблонов.

В главе 10 представлены результаты экспериментов. Отдельно оценивается качество извлечения ключевых терминов и качество автоматически полученных рефератов. Проведенные эксперименты подтверждают эффективность предложенных методов и алгоритмов.

В приложении 1 приведен фрагмент лингвистической базы знаний, содержащей дискурсивные маркеры и коннекторы, которые использовались в данном исследовании. В приложении 2 приведены примеры результатов работы созданной системы.

ГЛАВА 1

КЛАССИФИКАЦИЯ ПОДХОДОВ

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 авторефеирования — это преобразование, входными данными которого является текст (или несколько текстов), а результатом — аннотация, т. е. 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 (рис. 1):

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

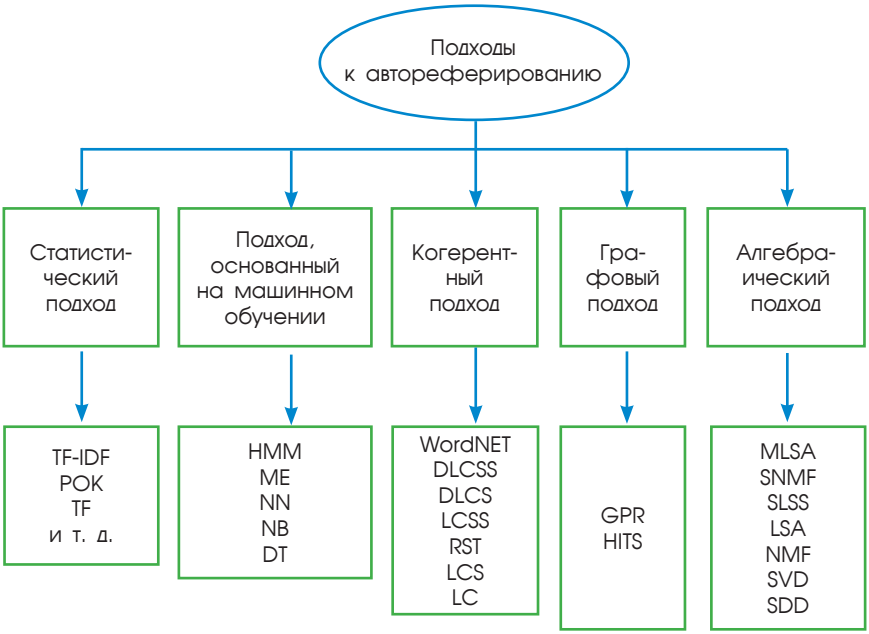

Рис. 1. Классификация подходов автореферирования текста

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. Сопряженными отношениями между элементами в тексте характеризуются ссылка, эллипсис, замещение, союз и лексическая когерентность, а также 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).

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. Кроме того, для обнаружения в тексте важных фрагментов 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 подход и подход, основанный на машинном обучении, — абстракцией. 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 табл. 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. В качестве входных данных 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 ранее оцененных запросов для поиска по веб-коллекции и коллекции нормативно-правовых документов

Окончание табл. 1

Наименования системы	Основные функции
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

Продолжение табл. 2

Наименование системы	Основные функции
Autonomy Knowledge Server	Анализ тектов 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 и ссылки, списки или кластеры
WebAnalyst	Представляет собой интеллектуальное масштабируемое клиент/серверное решение для компаний, желающих максимизировать эффект анализа данных в веб-среде. Сервер *WebAnalyst* функционирует как экспертная система сбора информации и управления контентом веб-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

Окончание табл. 2

Наименование системы	Основные функции
Intelligent Text Miner (IBM)	Технология эффективного анализа текстовых данных. Представляет собой набор отдельных утилит, запускаемых из командной строки, или скриптов, выполняемых 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 боль-

шей 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, а также создание на базе совмещенных алгоритмов новых, общедоступных сервисов интеллектуального поиска информации и систем автореферирования больших объёмов 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.

ГЛАВА 2

ЭКСТРАГИРУЮШИЕ МЕТОДЫ

Тексты рефератов, как уже говорилось ранее, могут полностью состоять из предложений, извлеченных из исходного текста (полная экстракция), и 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 стадия — 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 — формирование 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, расположению фрагмента в исходном тексте (заголовки, подзаголовки и т. д.), 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. Система *Open Text Summarizer* основана на похожих принципах [13]. Реферат составляется из предложений, содержащих слова с наибольшим весом. Для определения веса слов используются только статистические методы.

В работах Н. В. Лукашевич [14–16] 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 [17] 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 подсчитывается сумма левосторонних и правосторонних связей и происходит извлечение предложений с наибольшим весом. В статье [18] отмечается, что симметричное реферирование применимо не только к научным текстам, но и к новостным текстам различного объема. В существущей в открытом доступе системе *t-conspectus* [19] реализован принцип симметричного реферирования для определения значимости предложения. Для определения веса слов в этой системе используется мера *TF-IDF*.

Г. Эдмундсон [20] 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 [21–23] рассматривают задачу реферирования как задачу сокращения объема данных — исходный документ выступает в роли данных большой размерности, а задача реферирования — снизить размерность документа и сохранить его основное содержание. При использовании векторной модели для обработки текстовых данных получаются многомерные матрицы термы-на-документы, и требуется их редукция. В качестве методов редукции используется латентно-семантический анализ, в основе которого лежит сингулярное разложение матрицы.

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.

ГЛАВА 3

АБСТРАГИРУЮШИЕ 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 гг. 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 абза-

це и др. Авторы сопоставили предложения рефератов и предложения статей при помощи разработанной программы. Полученный корпус использовался для обучения алгоритма. Алгоритм был основан на наивном баейсовском классификаторе, который маркировал каждое предложение на включение или невключение его в реферат. 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 с целью их выборки 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). Лучшие резуль-

таты продемонстрировал 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*) для выбора перенесения в реферат тех или иных единиц.

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 — 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 указанного выше предложения. Такие связи называются риторическими отношениями.

В работе 1994 г. [29] предлагается вычислительная модель дискурса для японских информативных текстов, описывается 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 совре-

менной 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). С. А. Тревгода разработал систему, основанную на правилах вывода и узкоспециализированном словаре ключевых. Гибридный подход, предложенный П. Г. Осми-

ниным [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 [39, 40], в которых также формулируются утверждения об их свойствах. Работа [41] описывает опыт построения корпуса на русском языке, содержащего дискурсивные маркеры. Корпус общедоступный и включает в себя тексты разных жанров: научного, научно-популярного и новостного. Прежде чем использовать теорию риторических структур, приходится адаптировать ее для конкретного языка. Это связано с грамматическими языковыми 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 [42–44] 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.

ГЛАВА 4

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. Так, в системе *COMPENDIUM* [45] 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 стоится 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 [46]. 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* [47], которая генерирует рефераты для научно-технических документов. Авторы провели исследование в корпусе рефератов, написанных людьми, и выявили ряд трансформаций, которые применяли референты, например, слияние информации из различных частей документа, перефразирование оригинала и проч.

Подход авторов [48] 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 [49] 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* [50] и программу *MetaMap* [51] 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 отношениями 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 в качестве темы 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 [52] 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 [53, 54].

Авторы работы [55] предлагают подход к реферированию оценочных суждений или комментариев пользователей сети Интерне-

та. 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 [56] 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 и информация о структуре документа (самостоятельный или зависимый пункт формулы изобретения и т. д.). Реферированию патентов также посвящены работы [57–61].

К гибридному подходу могут быть отнесены графовые методы, в случае которых 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 [62, 63].

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.

В следующей главе обобщенно описана система *Scientific Text Summarizer*, реализующая предлагаемый нами гибридный метод автоматического построения аннотаций научных текстов.

ГЛАВА 5

SCIENTIFIC TEXT SUMMARIZER

Предположим, что есть текст T, очищенный после предварительной обработки. Он остоит из предложений $T = [s_1, \ldots, s_P]$.

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

$$\Psi : T \to \tilde{T}, |\tilde{T}| < |T|, |\tilde{T}| \leq 250 \text{ слов}.$$

Тогда алгоритм построения реферата можно описать в виде следующих 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. Аннотации и ключевые слова, написанные авторами, исключаются из текста и сохраняются отдельно, чтобы выполнять роль «золотого стандарта» и использоваться для оценки качества предложенного метода. Для лемматизации текста и построения морфологического словаря применяется программа *Mystem* [64]. 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 униграмм-

ная 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, и выполняются определенные преобразования с текстом (подробнее об этом в главе 9), в результате чего формируется квазиреферат: $T' = \left[s'_1, \ldots, s'_{P_1} \right], T' \subset T$.

4. Оценка весов предложений. 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. В итоге вес каждого предложения вычисляется по следующей формуле:

$$SW(s') = \frac{1}{L} \cdot \sum_{i=1}^{L} w_i + \frac{1}{M} \cdot \sum_{j=1}^{M} v_j + \frac{1}{N} \sum_{k=1}^{N} d_k,$$

где $W = \{w_1, \ldots, w_L\}$ — веса ключевых терминов ($|W| = L$). Веса w_i вычисляются как частоты ключевых слов (или многословных терминов) в тексте. $V = \{v_1, \ldots, v_M\}$ — веса значимых глаголов и существительных, которые часто встречаются в научных текстах ($|V| = M$). Веса v_j определяются из лингвистической базы знаний. $D = \{d_1, \ldots, d_N\}$ — веса дискурсивных маркеров и коннекторов $|D| = N$ Веса d_k определяются из лингвистической базы знаний.

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:

$$\tilde{T} = [s' \in T' : SW(s') > \beta]$$

где β = 0,15 является константой, которая определяется эмпирически. От нее зависит, насколько сильно будет сокращен текст.

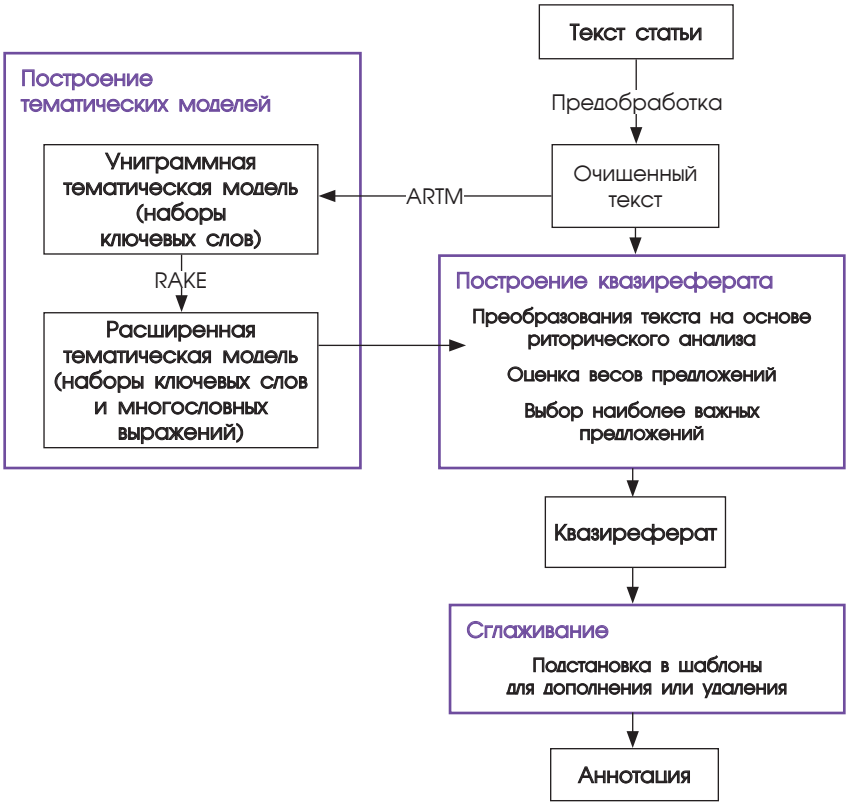

Рис. 2. Блок-схема системы *Scientific Text Summarizer*

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.

Предложенный гибридный метод реализован в системе *Scientific Text Summarizer* (рис. 2).

Система реализована на языке *Python 3* с использованием внешних библиотек *Scikitlearn, Gensim, NLTK, BigARTM, Flask и* некоторых других. Для поддержки созданной лингвистической базы знаний, описание которой приведено в главе 9 и приложении 1, использовалась *PostgreSQL.*

В следующих главах содержится более подробное описание кадого из этапов.

ГЛАВА 6

УНИГРАММНЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

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 [65]. 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 [66].

Тема — набор терминов (слов и словосочетаний), характеризующих принадлежность текста к определенной области знаний. Эти термины правильнее называть тематическими словами. С нашей точки зрения, их функции несколько отличаются 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, относя-

щаяся 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 $(w_1, ..., w_{n_d})$ 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 (рис. 3).

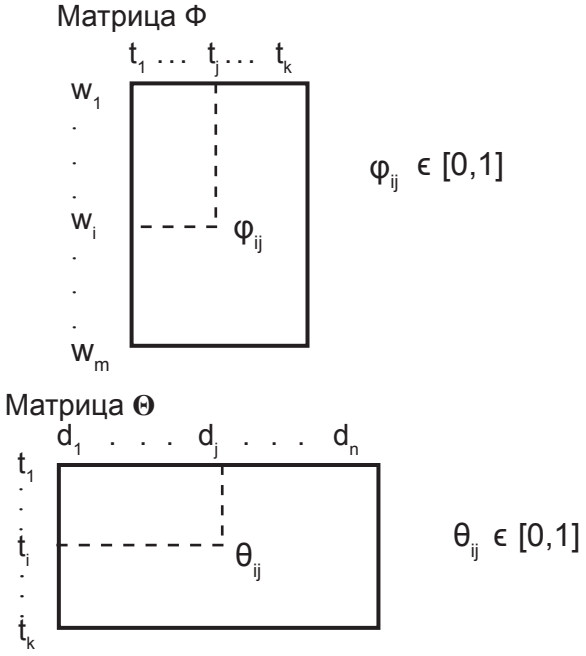

Рис. 3. Матрицы Φ и Θ

Гипотеза «мешка слов» состоит в том, что элементы выборки независимы, т. е. порядок слов в тексте документа не имеет значения, следовательно, 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_{d_w} 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 настоящее время тематические модели находят применение в самых различных областях [67]. Их используют для создания персонализированных медицинских рекомендаций, для библиографического анализа, анализа данных из социальных сетей, для многоязычного информационного поиска, для выявления трендов в новостных потоках или научных публикациях, 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 г. [68].

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 [69].

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 г. [70]. В настоящее время *ARTM* приобретает все большую популярность благодаря своей универсальности и гибкости настройки параметров моделей.

Современные требования к тематическим моделям довольно разнообразны. В основном они базируются на том, что тематические модели должны хорошо поддаваться интерпретации для лучшего понимания причин выделения определенных тем в тексте и структур самих тем конечному пользователю. 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.

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

Также для выбора базового алгоритма построения униграммных тематических моделей был проведен ряд экспериментов на части коллекции, подготовленной нами и описанной в главе 11. Описание этих экспериментов можно найти в работе [67].

Для оценки результатов были выбраны следующие метрики, реализованные в библиотеке *BigARTM* и описанные в работе [70]: перплексия, разреженность матриц Φ и Θ, доля фоновых слов, мощность ядер тем, чистота ядер тем, 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:

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

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**

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 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 $W_t = \{w \in W \mid p(t \mid w) > \delta\}$. **Мощность ядра темы** — это количество слов, содержащихся в лексическом ядре темы.

На основе ядра темы строятся следующие две оценки.

Чистота ядра темы — суммарная вероятность слов ядра:

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

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

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

На графике (рис. 4) представлена зависимость перплексии от числа итераций (проходов по коллекции).

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 в табл. 4.

По результатам эксперимента, приведенным в табл. 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. В совокупности с особенностями алгоритмов, представленными в табл. 3, было принято решение использовать алгоритм *ARTM* в реализации библиотеки *BigARTM* [66].

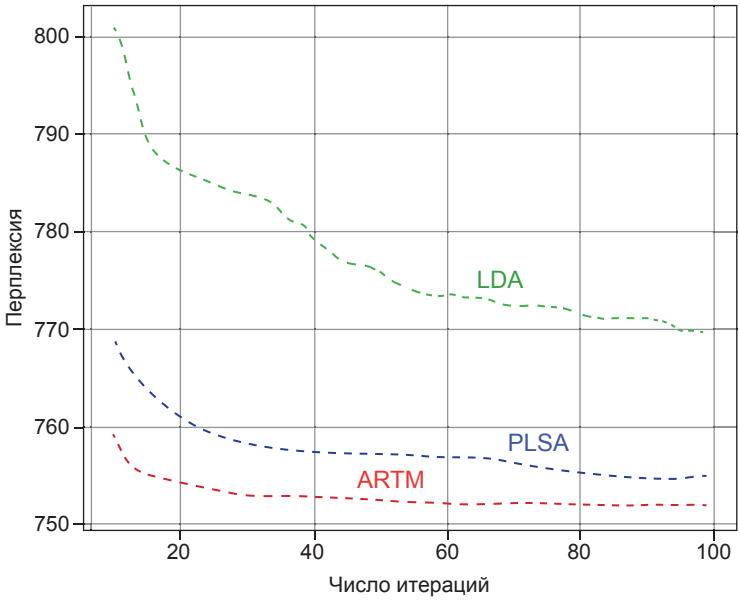

Рис. 4. Сравнение перплексии *LDA*, *PLSA* и *ARTM*

Таблица 4

Сравнение алгоритмов *PLSA* и *ARTM*

Метрика	***PLSA***	***ARTM***
Перплексия	754,784	751,888
Разреженность матрицы Φ	0,769	0,769
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

Он позволяет комбинировать тематические модели с помощью регуляризаторов. 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. *ARTM* гарантирует единственность и устойчивость решения; у него не наблюдается увеличения количества параметров модели с ростом числа документов, поэтому он может применяться к большим наборам данных. Кроме того, предложенная нами модификация (подробнее об этом в главе 7) позволяет использовать не только однословные, но и многословные выражения, что, на наш взгляд, повышает интерпретируемость модели.

ГЛАВА 7

ПРОБЛЕМА МНОГОСЛОВНЫХ ТЕРМИНОВ

Как упоминалось ранее, 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, слово «документ», встречающееся достаточно часто в настоящей работе, 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, например, в английском или немецком. Это связано с тем, что русский язык — флективный, т. е. язык, в котором 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. Так, если в тексте встретятся такие слово-

сочетания, 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. Так, для распознавания словоформ можно использовать словари, содержащие для каждого слова все его возможные формы [71]. 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. В работе [72] 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, 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 представленной 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 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:

9) n подряд идущих прилагательных ($n > 0$) + существительное в начальной форме;
10) существительное в начальной форме + n подряд идущих существительных в родительном падеже ($n > 0$).

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 $x, x_1, x_2, \ldots, x_i, \ldots, x_n$ — множество прилагательных из V; $y, y_1, y_2, \ldots, y_i, \ldots, y_m$ — множество существительных из V. Для морфологических признаков введем следующие обозначения: $z_1 = \{mal, fem, neu\}$ — содержит информацию о категории рода (мужской, женский, средний); $z_2 = \{sin, plu\}$ — содержит информацию о категории числа (единственное, множественное); $z_3 = \{nom, gen, dat, acc, ins, pre\}$ — 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:

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

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

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

Для извлечения многословных терминов из текстов используется алгоритм извлечения ключевых слов *RAKE (Rapid Automatic Keyword Extraction)*. Суть алгоритма описана в работе [73] и обобщенно состоит в следующем. На вход алгоритму подается текст, в котором алгоритм обнаруживает слова-кандидаты, представляющие собой отдельно стоящие содержательные слова или последовательности таких слов. Для отдельно стоящих слов-кандидатов вычисляется вес по формуле:

$$Score\left(w\right) = \frac{Deg\left(w\right)}{Freq\left(w\right)},$$

где $Deg\left(w\right)$ — степень слова в графе слов текста; $Freq\left(w\right)$ — частота слова в тексте.

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

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

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 [74].

Мера *TF-IDF* является произведением двух множителей:

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

$$tf\left(t,d\right)=\frac{n_t}{\sum\limits_k n_k},$$

где n_t — число вхождений слова t в документ d, $\sum\limits_k n_k$ — общее число слов в документе.

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

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

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

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

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

ГЛАВА 8

РАСШИРЕННЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ

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 рис. 5. Далее приведено подробное описание процесса построения расширенной тематической модели и извлечения ключевых фраз.

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 стро-

ится 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.

Рис. 5. Схема работы
модуля построения тематических моделей

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 подобрать как 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 расширить набор шаблонов или воспользоваться дополнительными способами согласования слов.

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.

Похожим 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.

ГЛАВА 9

РИТОРИЧЕСКИЙ АНАЛИЗ И ПРЕОБРАЗОВАНИЯ ГРАФОВ

В работе мы уже упоминали о том, что в теории риторических структур можно определить два типа ЭДЕ: ядро и сателлит. Ядро считается наиболее важной частью высказывания, так как содержит основную информацию. Сателлит поясняет ядро, содержит дополнительную информацию о нем и считается вторичным. Между тем выражения, в которых сателлит удален, могут быть поняты лишь в некоторой степени (рис. 6).

Маркеры (дискурсивные маркеры) — это слова или фразы, которые не имеют реального лексического значения, но вместо этого обладают важной функцией формирования разговорной структуры, передавая намерения говорящих при разговоре.

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

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, таким образом, чтобы получился сокращенный текст, являющийся промежуточным между исходным текстом и готовой аннотацией.

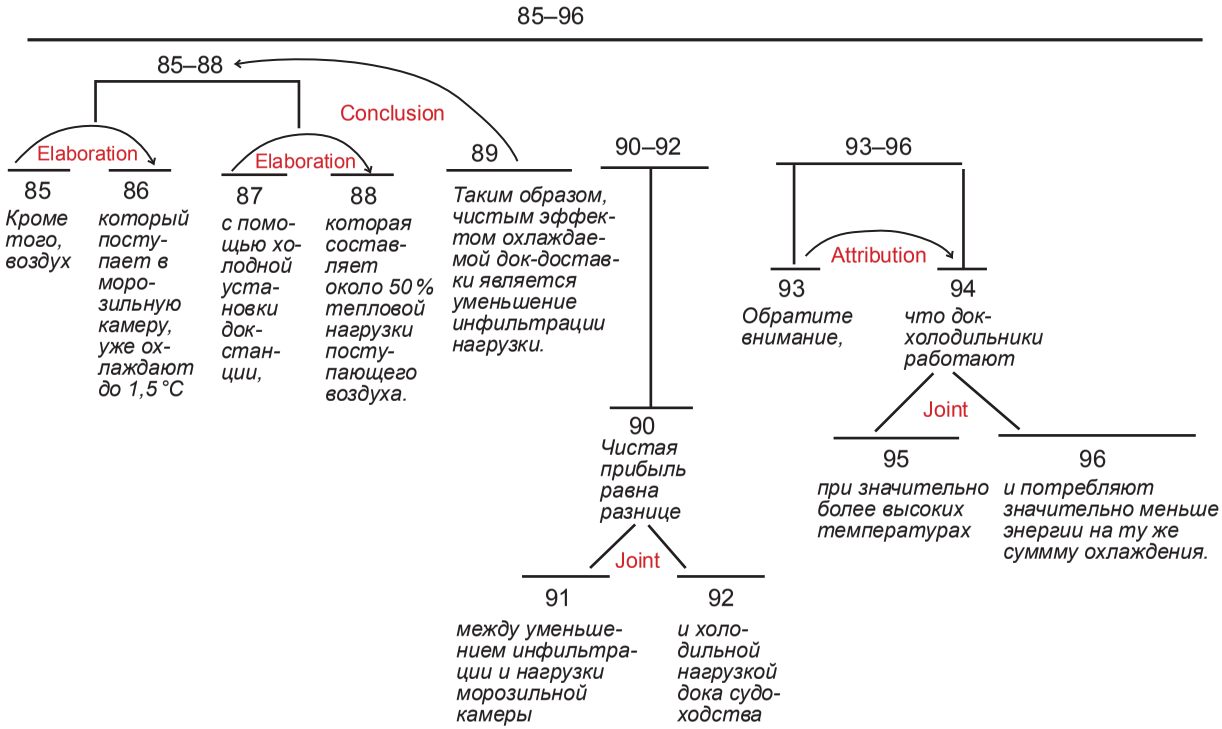

Рис. 6. Пример риторического анализа фрагмента текста

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 предыдущей главе, для рассмотренного примера действия, выполняемые системой на этом этапе, могут быть записаны в виде

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

Следовательно, сначала надо найти маркер y = «*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 табл. 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

За 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 из них.

MDELETE_SAVE: Это действие удаляет маркер предстоящего предложения и сохраняет предложение с заданным маркером (рис. 7).

MDELETE_SAVE: Конечный результат после операций (рис. 8).

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 охлажда-*

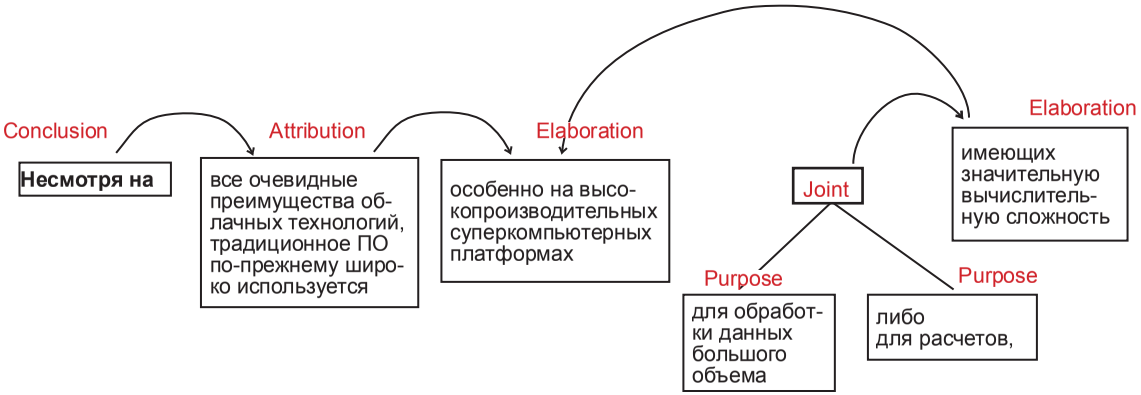

Рис. 7. Пояснение действия MDELETE_SAVE

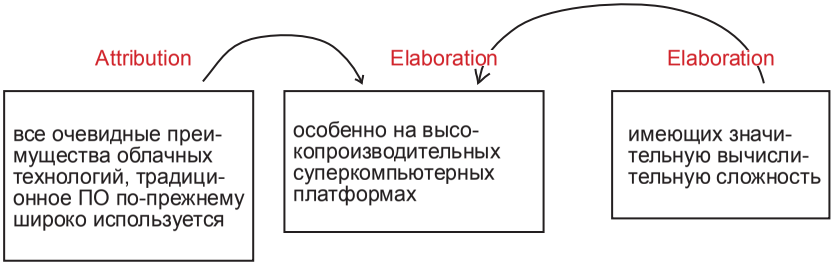

Рис. 8. Результат MDELETE_SAVE

емой 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:

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

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 квазиреферата 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.

ГЛАВА 10

СГЛАЖИВАНИЕ ТЕКСТА АННОТАЦИИ

Как уже отмечалось ранее, 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 словосочетани, как показано в табл. 6.

Таблица 6

Результат применения сглаживания

№	До сглаживания	После сглаживания
1	Рассматриваем	Рассмотрено
2	Можно сделать вывод	Сделан вывод
3	Для этого предлагается	Предложено
4	Были апробированы	Апробировано
5	В статье приведен	Преведен
6	На основе данных рассуждений создан	Создан
7	В нашей статье	В статье
8	Важной составляющей	Составляющей
9	Основной идеей	Основой
10	Как показывает практика	–
11	Как показали эксперименты	–
12	Целью данной работы	Целью работы
13	В статье рассмотрены	Рассмотрены
14	Данная операция	Операция

Для сглаживания предложений используются шаблоны двух видов: 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 подходоящих 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. $X \in$ {«В статье», «В работе», …};

Y_v — глагол $V \in$ {«рассматриваются», «рассматривается», …};

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

Z — оставшаяся часть предложения (сателлит).

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

X — добавляемый фрагмент. $X \in$ {«Новизна», «Новизна и перспективность», …};

Y_n — существительное $N \in$ {«метода», «алгоритма», …};

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

Z — оставшаяся часть предложения (сателлит).

Шаблон «Цель» представлен несколькими вариантами.

Вариант 1

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).

Вариант 2

$Y_V \in$ {Показана, Представлена, Исследуется, …};

$Y_N \in$ {целесообразность взаимодействия};

KW — ключевые слова;

$T \in$ {с системой, на основе, по вопросам, …};

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

Вариант 3

$Y_N \in$ {Применение, Использование, Разработка, Вычисление, …};

$X_w \in$ {данной работы, статьи, исследования, модели, этого, …};

$K_p \in$ {полезно для};

KW — ключевые слова;

$P \in$ {с целью, …};

$P_p \in$ {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

$Y_V \in$ {Рассматриваются, Проводятся, Перечислены, Предлагаются, …};

$Y_N \in$ {методы, методика, Система, возможности, задачи, … };

$O \in$ {где основой являются, где используются, …};

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

Вариант 2

$X \in$ {В статье, В данной работе, В данной статье, В модели, В информационных системах, …};

$\in$ {Рассматриваются, Проводятся, Перечислены, Предлагаются, …};

KW — ключевые слова;

$T \in$ {где применяются, с применением, …};

KW — ключевые слова;

$K_d \in$ {каждое из которых, которые, примером являются, с применением, …};

KW — ключевые слова.

Вариант 3

$X \in$ {Создание, Применение, Использование, Разработка, Вычисление, …};

$Y_N \in$ {метода, методики, системы, …};

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

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

«Реализация»

Вариант 1

$Y_N \in$ {Алгоритм, системы, …};

$Y_V \in$ {реализован, реализованы, …};

$PREP \in$ {на, в, …};

$KW \in$ {языке C++, …}.

Вариант 2

$Y_V \in$ {Описана, …};

$Y_N \in$ {Программная реализация, программное обеспечение, …};

$K_r \in$ {разработанного алгоритма, …};

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

«Недостатки»

$Y_N \in$ {Недостаток, Достоинства, …};

$N \in$ {методов, …};

PREP $\in$ {в том, что; …};

$Y_V \in$ {рассматривают, …};

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

«Заключение»

$Y_V \in$ {Приведены, Рассмотрены, …};

KW — ключевые слова;

$K_C \in$ {таким образом, чтобы; где можно сделать вывод, …};

Z — оставшаяся часть предложения (сателлит).

ГЛАВА 11

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для проведения экспериментов мы собрали коллекцию научных статей на русском языке из находящихся в открытом доступе материалов международного научно-практического журнала «Программные продукты и системы» за 2010–2018 гг. Объем этой коллекции составляет 1200 статей.

Далее в этой главе приведены несколько примеров работы алгоритма для различных документов разных тематик. В табл. 7 представлены некоторые из наиболее частотных терминов для первых трех тем расширенной тематической модели коллекции.

По представленным в таблице результатам можно отметить, что граница между узкими темами не слишком четкая: если тема 3 довольно хорошо интерпретируема как отдельная предметная область, связанная с управлением проектами и процессом разработки программного обеспечения, темы 1 и 2 обе связаны с классификацией и распознаванием.

В табл. 8 представлены извлеченные программой ключевые термины для нескольких научных публикаций.

Таблица 7

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	'система', 'управление', 'процесс', 'модель', 'требование', 'разработка', 'система управления', 'орган управления', 'процесс разработки', 'модель прогнозирования', 'критерий эффективности проекта', 'этап прогнозирования', 'критерий эффективности', 'эффективность проекта'

Таблица 8

Примеры ключевых терминов

№	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 интеллекта	программный, пользователь, система управления, уровень развития, нечеткий, интерфейс, характеристика, эксперт, система управления
3	Прогнозирование платежеспособности клиентов банка на основе методов машинного обучения и марковских цепей	прогнозирование, состояние, клиент, классификатор, заемщик, решение задачи, дерево решения
4	Разработка системы хранения ансамблей нейросетевых моделей	данные, модель, набор, ансамбль, ряд, преобразование, хранение, нейросетевой, оценка качества, процесс формирования, классификация текста

Ключевые термины, извлеченные автоматически при помощи *Scientific Text Summarizer*, 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, термин «дерево решения» дает понять, что в качестве алгоритма машинного обучения в статье 3 использовались деревья решений, а термин «классификация текста» в статье 4 указывает, что ансамбли нейросетевых моделей здесь использовались для классификации текста (а не только, например, для классификации изображений).

Оценка качества извлечения ключевых слов

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

Точность — показатель количества правильных положительных решений:

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

Полнота — показатель того, сколько всего ключевых слов найдено:

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

F-мера –гармоническое среднее между точностью и полнотой:

$$F_{measure} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

В табл. 9 содержится сравнение качества извлечения ключевых терминов разрботанной системой (*Scientific Text Summarizer*) и другими системами, находящимися в открытом доступе в сети Интернет. Сравнение проводилось на 260 текстах нашей коллекции.

Таблица 9

Оценка качества извлечения ключевых терминов

Система	Метод	Точность, %	Полнота, %	F-мера, %
t-conspectus (2015)	*TF-IDF,* симметричное реферирование	4	7	5
Open Text Summarizer (2016)	Статистический	3	25	5
***Scientific Text Summarizer* (2018)**	Комбинированный	**44**	**37**	**40**

Для того чтобы понять, насколько хорошо наша система определяет ключевые слова по тексту научной статьи, мы смотрели на совпадение выделенных автоматически ключевых слов с теми, которые авторы указывают сами (когда статья опубликована, после аннотации приводится список ключевых слов; авторы выписывают их вручную). Нами было замечено, что те слова, которые были указаны авторами, иногда вообще не встречаются в тексте статьи, так как среди них могут быть общие термины. Например, обозначается область исследований, а в статье приводятся более конкретные рассуждения; общее название какого-то метода, в то время как в статье рассматривается и называется только более конкретная разновидность этого метода; приведена аббревиатура, а в тексте статьи используется полное название (или наоборот). Подобные несоответствия при автоматической оценке относятся к ошибке, хотя такие термины в данном случае считаются семантически эквивалентными. Из чего можно сделать вывод о том, что точность и полнота, являющиеся общепризнанными мерами качества, имеют недостаток, который заключается в отсутствии возможности учитывать подобные случаи, что, по нашему мнению, является одной из причин относительно низких значений точности и полноты (см. табл. 9).

Оценка качества авторефератов

Согласно [75], пока не существует общепринятого эффективного способа автоматической оценки систем автореферирования, поэтому мы оценивали результаты, основываясь на экспертной оценке, а также автоматически при помощи точности, полноты и *F*-меры.

Точность полученных аннотаций, оцененная экспертами, оказалась довольно высокой. Экспертная оценка результатов реферирования показала, что 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них (что 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].

В ходе эксперимента мы заметили, что авторы нередко 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 со стилем изложения самой статьи и чаще всего означает, что в тексте имеется много содержательных предложений.

Точность, полноту и *F*-меру мы вычисляли способом, похожим на [30, 37]. 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:

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

Сравнительная оценка результатов автореферирования приведена в табл. 10.

Таблица 10

Оценка результатов автореферирования

Система	Метод	Точность, %	Полнота, %	F-мера, %
Коллекция текстов на русском языке				
Trevgoda (2009)	Шаблоны, риторический анализ	67,03	64,81	66,03
t-conspectus (2015)	*TF-IDF*, симметричное реферирование	10,00	22,20	36,30
Open Text Summarizer (2016)	Статистический	12,00	24,20	38,50
***Scientific Text Summarizer* (2018)**	Комбинированный	**75,23**	**68,21**	**71,55**
Коллекция текстов на английском языке				
Marcu (1998)	Комбинация эвристик	73,53	67,57	70,42

Результаты *Marcu* (1998) и *Trevgoda* (2009) были взяты из научных работ, опубликованных в открытом доступе [30, 37], поэтому следует учитывать, что авторы проводили эксперименты на других коллекциях, отличных от нашей. Результаты работы систем *t-conspectus* [19] и *Open Text Summarizer* [13] были получены на части нашей коллекции (260 документов). Из таблицы видно, что полученные результаты подтверждают эффективность предложенных методов.

В прил. 2 приведены примеры промежуточных и конечных результатов работы системы *Scientific Text Summarizer*. Возможное улучшение предложенного метода, по нашему мнению, состоит в том, чтобы дополнить правила удаления менее важных предложений, увеличить количество шаблонов для сглаживания, расширить списки маркеров и коннекторов.

ЗАКЛЮЧЕНИЕ

Основной целью исследований, изложенных в монографии, является создание нового оригинального метода автоматического реферирования научно-технических текстов. Разработанный гибридный метод позволяет получать рефераты (аннотации) высокого качества и определять темы текстов в виде набора ключевых терминов, используя методы тематического моделирования и машинного обучения, графовое представление текстов и риторической анализ.

Формально описанная методика обнаружения важных элементов в тексте базируется на понятиях теории риторических структур. В ходе исследования была создана лингвистическая база знаний на основе анализа подъязыка рефератов, используемая для оценки весов предложений квазиреферата. Предложен алгоритм построения расширенных тематических моделей коллекций текстовых документов и описана процедура сглаживания предложений, позволяющая сделать текст полученного реферата (аннотации) более связным и последовательным.

Представленные модели и алгоритмы реализованы в виде системы, позволяющей автоматически формировать аннотации статей научно-технической тематики. 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 результатов будет увеличено количество шаблонов для сглаживания, дополнены списки маркеров и коннекторов. Запланированы 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

Автореферирование — это составление коротких изложений материалов, аннотаций или рефератов, 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).

APTM — 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 документа.

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.

СПИСОК ЛИТЕРАТУРЫ

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. 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 : тр. Междунар. конф. «Диалог-2004» // Компьютерная лингвистика и интеллектуальные технологии. М. : Наука, 2004. С. 579–591.

4. *Kupiec J., Pederson J., Chen F.* A trainable document summarizer // In Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval. Seattle, 1995. P. 68–73.

5. *Танатар Н. В., Федорчук А. Г.* Интеллектуальные поисково-аналитические системы мониторинга СМИ : науч.-практич. и теор. сборник. Киев, 2008. 477 с.

6. *Михаилян А.* Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах. 2000. URL: http://citforum.ru/programming/digest/avtestlang.shtml (дата обращения: 11.10.2018).

7. *Харламов А. А.* Автоматический структурный анализ текстов // Открытые системы. М., 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. URL: http://www.rco.ru/?page_id=4875 (дата обращения: 11.10.2018).

10. *Бурмистров А. С., Свиридова О. В.* Экспертая оценка программных продуктов для аннотирования документов // Постулат. 2017. № 5. 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. и др.* 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. *Andonov F., Slavova V., Petrov G.* On the Open Text Summarizer // International Journal "Information Content and Processing". 2016. Vol. 3. № 3. URL: http://www.foibg.com/ijicp/vol03/ijicp03-03-p05.pdf (дата обращения: 12.11.2019).

14. *Лукашевич Н. В.* Автоматическое построение аннотаций на основе тематического представления текста : тр. Междунар. семинара «Диалог-1997». М., 1997. С. 188–191.

15. *Лукашевич Н. В., Добров Б. В.* Построение структурной тематической аннотации текста : тр. Междунар. семинара «Диалог-1998». М., 1998. Т. 2. С. 795–802.

16. *Лукашевич Н. В., Добров Б. В.* Автоматическое аннотирование новостного кластера на основе тематического представления : тр. Междунар. конф. «Диалог-2009» // Компьютерная лингвистика и интеллектуальные технологии. М. : Изд-Во РГГУ, 2009. Т. 8. С. 27–31.

17. *Яцко В. А.* Симметричное реферирование: теоретические основы и методика // НТИ. Серия 2. Информационные процессы и системы. 2002. № 5. С. 18–28.

18. *Вичева О. Н.* Подходы к автоматическому обзорному реферированию группы текстов одной тематики : сб. науч. статей «Проблемы современной прикладной лингвистики». Минск : МГЛУ, 2014. С. 246–252.

19. *Butakov A.* T-CONSPECTUS. 2015 URL: http://tconspectus.pythonanywhere.com/about#algorithm (дата обращения: 12.11.2019).

20. *Edmundson H. P.* New methods in automatic extracting // Journal of the ACM (JACM). 1969. Vol. 16. № 2. P. 264–285.

21. Automatic Text Summarization Using Latent Semantic Analysis // Programming and Computer Software. 2011. Vol. 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, 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. 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. Vol. 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». 2016. 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). P. 409–445.

34. *Bosma W.* Query-Based Summarization using Rhetorical Structure Theory // 15th Meeting of CLIN. 2015. P. 29–44.

35. *Huspi S.* H. Improving Single Document Summarization in a Multi-Document Environment : PhD thesis. Melbourne, Australia : RMIT University, 2017. 190 p.

36. *Mithun S.* Exploiting rhetorical relations in blog summarization : PhD thesis, Montreal, Canada : Concordia University, 2010. 230 p.

37. *Тревгода С. А.* Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений : автореф. дис. … канд. тех. наук (05.13.01) / С. А. Тревгода. СПб. : Санкт-Петербургский гос. электротехн. ун-т, 2009. 15 с.

38. *Осминин П. Г*. Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод : дис. ... канд. филол. наук (10.02.21) / П. Г. Осминин. Челябинск : Южно-Уральский гос. ун-т, 2016. 239 с.

39. *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.

40. *Bakiyeva A. M., Batura T. V., Yerimbetova A. S. et al.* Methods for constructing natural language analyzers based on Link Grammar and rhetorical structure theory // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2016. Is. 40. P. 37–51.

41. *Pisarevskaya D., Ananyeva M., Kobozeva M. et al.* Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. 2017. Iss. 16 (23). Vol. 1. P. 194–204.

42. *Khan A., Salim N., Kumar Y. A* Framework for multi-document abstractive summarization based on semantic role labelling // Applied Soft Computing. 2015. Vol. 30. P. 737–747.

43. *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.

44. *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. P. 64–73.

45. *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. P. 164–175.

46. *Hovy E., Lin Ch.-Y.* Automated text summarization and the SUMMARIST system // Proceedings of the TIPSTER Text Program. 1998. P. 197–214.

47. *Saggion H., Lapalme G*. Generating indicative-informative summaries with SumUM // Computational Linguistics. 2002. Vol. 28. № 4. P. 497–526.

48. *Foster G. F*. Statistical lexical disambiguation: Master's thesis. Montreal, Canada : McGill University, School of Computer Science, 1991. 340 p.

49. *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.

50. Unified Medical Language System (UMLS). 2016. URL: http://www.nlm.nih.gov/research/umls/ (дата обращения: 11.10.2018).

51. *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.

52. *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.

53. *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.

54. *Megala S., Kavitha A., Marimuthu A*. Feature Extraction Based Legal Document Summarization // International Journal of Advance Research in Computer Science and Management Studies. 2014. Vol. 2. Iss. 12. P. 346–352.

55. *Lloret E., Boldrini E., Vodolazova T. et al.* A novel concept-level approach for ultra-concise opinion summarization // Expert Systems with Applications. 2015. Vol. 42, Iss. 20. P. 7148–7156.

56. *Brügmann S., Bouayad-Aghab N., Burga A. et al.* Towards content-oriented patent document processing: Intelligent patent analysis and summarization // World Patent Information. 2015. Vol. 40. P. 30–42.

57. *Mahdabi P., Andersson L., Hanbury A., Crestani F.* Report on the CLEF-IP 2011. Experiments: Exploring Patent Summarization. 2011. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.7897&rep=rep1&type=pdf (дата обращения: 11.10.2018).

58. *Wanner L.* Generation of Patent Abstracts: A Challenge for Automatic Text Summarization // Proceedings of the SEPLN 2012 workshops: E-LKR and ATSF. 2012. URL: http://ceur-ws.org/Vol-882/elkr_atsf_2012_keynote.pdf (дата обращения: 11.10.2018).

59. *Chieze E.* An Automatic System for Summarization and Information Extraction of Legal Information // Semantic Processing of Legal Texts / E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia. Heidelberg, 2010. P. 216–234.

60. *Goldstein A*. Generation of Natural-Language Textual Summaries from Longitudinal Clinical Records // Studies in Health Technology and Informatics. 2015. Vol. 216. P. 594–598.

61. *Goldstein A*. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data // Journal of Biomedical Informatics. 2016. Vol. 61. P. 159–175.

62. *Анисимов А. В., Марченко А. А.* Ассоциативное реферирование естественно-языковых текстов // Штучний інтелект. 2006. № 3. С. 488–492.

63. *Попов М. Ю., Заболеева-Зотова А. В., Фоменков С. А.* Визуализация семантической структуры и реферирование текстов на естественном языке. 2003. URL: http://www.dialog-21.ru/media/2725/popov.pdf (дата обращения: 11.10.2018).

64. *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. P. 273–280.

65. *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке : тр. Ин-та сист. программирования РАН. 2012. С. 215–242.

66. *Воронцов К. В., Фрей А. И., Апишев М. А. и др.* BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций : XVII Междунар. конф. DAMDID/RCDL'2015 // Аналитика и управление данными в областях с интенсивным использованием данных. Обнинск, 2015. URL: http://www.machinelearning.ru/wiki/images/e/e4/Voron15damdid.pdf (дата обращения: 06.12.2018)

67. *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.

68. *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. P. 289–296.

69. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. № 3. P. 993–1022.

70. *Воронцов К. В., Потапенко А. А.* Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем : тр. Междунар. конф. «Диалог-2014» // Компьютерная лингвистика и интеллектуальные технологии. Вып. 13(20). М. : Изд-во РГГУ, 2014. С. 676-687.

71. *Кипяткова И. С., Карпов А. А.* Аналитический обзор систем распознавания русской речи с большим словарем : тр. СПИИРАН, 2010. Т. 12. С. 7–20.

72. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А. и др.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста : тр. Междунар. конф. «Диалог-2007» // Компьютерная лингвистика и интеллектуальные технологии. М. : Изд-во РГГУ, 2007. С. 70–75.

73. *Rose S., Engel D., Cramer N., Cowley W.* Automatic keyword extraction from individual documents // Text Mining: Applications and Theory. 2010. P. 3–20.

74. *Leskovec J., Rajaraman A., Ullman J. D.* Mining of Massive Datasets. Cambridge, UK : Cambridge University Press, 2014. 476 p.

75. *Das D., Martins A. A.* Survey on Automatic Text Summarization : Technical report // Literature Survey for the Language and Statistics II course at Carnegie Mellon University. Pittsburgh, US. 2007. P. 192–195.

ПРИЛОЖЕНИЕ 1

МАРКЕРЫ И КОННЕКТОРЫ

Таблица П1

Некоторые риторические маркеры

Elaboration (Детализация) *El1*: Вследствие (того, чего), *El2*: Кроме того, *EL3*: Например, *EL4*: в том числе, *El5*: В частности *El6*: Их можно условно разбить\ разделить	**Concession (Уступка)** *Conc1*: Поскольку, *Conc2*: Исходя из этого, *Conc3*: Хотя,
Restatement (Переформулировка) *Res1*: То есть, *Res2*: Иными словами, *Res3*: Иначе говоря *Res3*: Между тем, практика показывает	**Contrast (Контраст)** *Cont1*: Однако, *Cont2*: Несмотря на то, что *Cont3*: Обратите внимание Cont4: Но
Purpose (Цель) *Pur1*: Для того, чтобы *Pur2*: Чтобы *Pur3*: целью которого *Pur4*: целью чего	**Evidence (Обоснование)** *Ev1*: Очевидно, что *Ev2*: Доказательством тому, *Ev3*: Доказательство чего, *Ev4*: Таким образом, *Ev5*: Безусловно, *Ev6*: Можно сделать вывод

	Ev7: Как показала (практика), *Ev8*: Основной же проблемой *Ev9*: Важной составляющей *Ev10*: Преимуществами системы *Ev11*: В трактовке *Ev12*: Исследования показали *Ev13*: Основной идеей *Ev14*: В настоящей работе *Ev15*: В данной работе *Ev16*: В данной статье *Ev17*: На практике приведенный алгоритм *Ev18*: Важное преимущество
Cause-Effect (Причина) *CEf1*: Почему *CEf2*: Из-за *CEf3*: Так как *CEf4*: Поэтому *CEf5*: Потому	**Background (Фон)** *Bg1:* При этом *Bg2:* При том *Bg3:* Для его внедрения/использования

Таблица П2

Краткая таблица маркеров и действий

№	Название маркера	Маркер	Действие
1	Background	При этом	save_delete
2	Cause-Effect	Почему	Mdelete_save
3	Cause-Effect	Из-за	Mdelete_save
4	Cause-Effect	Так как	Mdelete_save
5	Cause-Effect	Поэтому	save_save
6	Cause-Effect	Потому	Mdelete_save
7	Cause-Effect	поэтому	mdelete_save
8	Comparison	Больше чем,	save_delete
9	Concession	Поскольку	delete_save
10	Concession	Исходя из этого	mdelete_save
11	Concession	Хотя	Mdelete_save

Продолжение табл. П2

№	Название маркера	Маркер	Действие
12	Concession	Как видно из таблицы	delete_save
13	Concession	Как видно из рисунка	delete_save
14	Concession	Соответсвенно	delete_save
15	Concession	Наконец	delete_save
16	Concession	В заключение отметим	save
17	Concession	В результате работы	save
18	Contrast	Однако	save_save
19	Contrast	Несмотря на то, что	save_delete
20	Contrast	Обратите внимание	save_delete
21	Contrast	однако	save
22	Contrast	несмотря на то, что	save_mdelete
23	Elaboration	например, о том	delete_example
24	Elaboration	Например	save_delete
25	Elaboration	Вследствие того	save_delete
26	Elaboration	Вследствие чего	save_delete
27	Elaboration	Кроме того	save_delete
28	Elaboration	В том числе	save_delete
29	Elaboration	В частности	save_delete
30	Elaboration	например	delete_example
31	Elaboration	кроме того	save_mdelete
32	Elaboration	к примеру	save_mdelete
33	Elaboration	Необходимо отметить	save_delete
34	Elaboration	В связи с этим	save_delete
35	Elaboration	в том числе	delete_example
36	Elaboration	а именно	delete_example
37	Elaboration	Кроме этого	save_save"
38	Evidence	Таким образом	mdelete_save
39	Evidence	Очевидно, что	mdelete_save
40	Evidence	В данной работе	save
41	Evidence	В статье	save
42	Purpose	Для того, чтобы	delete_save
43	Purpose	Чтобы	save

Окончание табл. П2

№	Название маркера	Маркер	Действие
44	Restatement	То есть	save_delete
45	Restatement	Иными словами	save_delete
46	Restatement	Иначе говоря	save_delete
47	Restatement	Речь идет	save_delete
48	Restatement	Другими словами	save_delete
49	Restatement	Соответственно	save_delete
50	Restatement	В то же время	save_delete

Таблица П3

Примеры маркеров и действий

№	Маркер	Действие	Примеры
1	Вследствие (того, чего)	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. ***Вследствие*** *этого сокращаются сроки разработки прикладного программного обеспечения и проведения вычислительных экспериментов*
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. ***Кроме того****, в предлагаемой модели поддерживается возможность сохранения в клетке следов пребывания агентов*
3	Например, например	Save_delete	В соответствии с методом по формулам (3) для каждого критерия вычисляются значения базового распределения доверия. ***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(Θ)=0,208.*

№	Маркер	Действие	Примеры
4	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.
5	В частности в частности	Save_mdelete	Если затраты на производственные ресурсы снизить достаточно сложно, то затраты на обеспечение безопасных условий труда можно значительно сократить, ***в частности****, за счет проектирования новых рабочих мест, на которых исключено нарушение техники безопасности*
6	Их можно условно разбить/разделить Их можно разбить/разделить	Save_save	Неопределенный характер температурных полей технических систем обусловлен неопределенным характером факторов, определяющих тепловой режим технической системы. ***Их можно условно*** разбить на три группы: факторы конструкции технической системы, факторы, возникающие при функционировании технической системы, и факторы окружающей среды
7	Поскольку	Mdelete_save	**Поскольку** 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
8	Исходя из этого	Mdelete_save	***Исходя из этого,*** разработанная модель и программный комплекс могут применяться для моделирования процесса непрерывного литья цилиндрических заготовок из цветных металлов

№	Маркер	Действие	Примеры
9	Хотя	Save	**Хотя** использование ГЛОНАСС весьма актуально, анализ показывает, что технология позиционирования и идентификации мобильных объектов на пространственных цифровых моделях в транспортной сфере развита недостаточно
10	То есть	Save_delete	Хороших результатов работы метода удалось добиться на слипшихся клетках, относящихся к разным типам. *То есть слипшиеся лейкоциты и эритроциты (как, например, на рисунке 4) удается корректно разделить если не всегда (2 % ошибок приходятся как раз на случай слипшихся клеток), то в большинстве случаев*
11	Иными словами	Save_delete	Реальные температурные поля технических систем, как показывает практика, не являются строго определенными и детерминированными, а носят неопределенный, а точнее интервальный характер. ***Иными словами**, температура в каждой точке технической системы может принимать любые значения внутри некоторых интервалов своего изменения.* Как известно, для особого семейства МЦ, называемых эргодичными, по прошествии длительного периода времени вероятность попадания случайной величины в то или иное состояние перестает зависеть от начального состояния цепи. ***Иными словами**, при $i \to \infty$ $P(i)ab=P(i)b$.* Для данного случая это условие можно выразить так: возведенная в некоторую степень матрица перехода не содержит нулевых элементов. ***Иными словами**, у МЦ есть вероятность через определенное число шагов перейти из любого состояния в какое-либо другое*

№	Маркер	Действие	Примеры
12	иначе говоря	Save_delete	Существующие методы моделирования температурных полей технических систем исходят из допущения, что параметры, определяющие тепловые режимы, являются детерминированными, ***иначе говоря***, *все данные, определяющие протекание теплового процесса и его характер в технической системе, являются полностью известными и однозначно определенными.* Следующим модулем, в котором агрегируется совокупность действий одного уровня, последовательно реализуемых в рамках выполняемой процедуры, является шаг соответствующего уровня, ***иначе говоря***, *шаг — это неделимый (законченный) набор элементарных действий*
13	Однако однако	Delete Save Delete_msave	*Данные логики лишены недостатков с точки зрения однозначности формулируемых на их базе свойств.* ***Однако***, 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. 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*

№	Маркер	Действие	Примеры
14	Несмотря на то, что	Delete_save	Исторически типичным в таких случаях является решение *специальным ПО* (СПО) функциональных задач ИС как во взаимодействии с локальной БД, так и при вводимых оператором внешних данных и обмене специализированными сообщениями между СПО, размещенным в различных узлах ИС. ***Несмотря на то, что*** *значительная часть задач, решаемых в различных узлах ИС, идентична или подобна, обмен сообщениями между СПО существенно препятствует его унификации.* В этом случае для любого человека независимо от места его проживания откроется возможность получить образование мирового класса. *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].* 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]. *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*

№	Маркер	Действие	Примеры
15	Обратите внимание	Save_delete	Все этапы формирования онтологии вместе с ее оценкой можно свести к схеме, представленной на рисунке 2 [2]. *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*
16	Для того, чтобы	Save	Однако необходимы еще более глубокое осмысление получаемых результатов и дополнительные исследования **для того, чтобы** с помощью программного комплекса получать и анализировать действительно наиболее важную информацию
17	Для того, чтобы Чтобы	Save	**Чтобы** избежать переобучения, количество обучающих примеров должно быть соразмерно числу используемых терминов
18	целью которого	Save	Корнелльский университет реализовал проект «Matlab on the Teragrid» [1], **целью которого** являлось предоставление Matlab пользователям Teragrid в качестве сервиса, в том числе с использованием порталов научного взаимодействия, таких как nanohub.org [2]
19	Целью данной работы	Save	**Целью данной работы** является разработка веб-сервиса, автоматизирующего реализацию баз знаний продукционного типа на основе результатов концептуального (когнитивного) моделирования
20	Очевидно, что	Mdelete_Save	***Очевидно, что*** критическим аспектом приведенной классификации является соотношение операционной нагрузки и локальных вычислительных возможностей

№	Маркер	Действие	Примеры
21	Таким образом,	Mdelete_Save	***Таким образом,*** все сервисные операции с ЭБД выполняются автоматически, без участия экипажа
22	можно сделать вывод		*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,* ***можно сделать вывод***, что универсального метода решения задачи оптимального размещения не существует, в каждой конкретной задаче есть свои особенности и ограничения, которые необходимо учитывать
23	Можно сделать вывод	Save	***Можно сделать вывод*** о сложности данной темы и необходимости усовершенствовать преподнесение материала в рамках семинарских занятий
24	Как показала/показывает практика Как показали эксперименты	Save	***Как показала практика***, таким инструментом может быть простая таблица, содержащая два столбца: в одном указываются задачи ТЗ, в другом — соответствующие им прецеденты (табл. 1).
25	Важной составляющей	Save	***Важной составляющей*** имитационно-тренажерных комплексов является система управления
26	Преимуществами системы	Save	**Преимуществами системы** являются простота ее использования, нетребовательность к ресурсам и расширяемость
27	В трактовке	Mdelete_Save	***В трактовке*** стандарта POSIX-2001 в трассировке логически участвуют три процесса, которые физически могут совпадать между собой: трассируемый (целевой), трассирующий (управляющий трассировкой) и анализирующий данные трассировки

№	Маркер	Действие	Примеры
28	Исследования показали	Save	***Исследования показали***, что наилучший результат получается при удалении всей иерархии внутри блока перед синтезом
29	Основной идеей	Save	***Основной идеей*** технологии кеинга является выделение объекта от однородного фона.
30	В настоящей работе	Save	***В настоящей работе*** развивается метод математического и компьютерного моделирования интервально стохастических температурных полей, обусловленных интервальным стохастическим характером входных данных, определяющих тепловые режимы технической системы
31	В данной работе	Save	***В данной работе*** 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]
32	В данной статье	Save	***В данной статье*** описывается реализация генетического алгоритма для выявления и отбора наиболее релевантных результатов, полученных в ходе последовательно выполняемых операций тематического поиска
33	На практике приведен	Save	На практике приведенный алгоритм необходимо модифицировать прокладкой перекрестных маршрутов между всеми *процессорными элементами* (ПЭ) и ограничениями на просмотр портов коммутаторов (отдельные крэйты в сложной системе могут включаться неодновременно) — необходима локализация алгоритма в крэйте или в группе крэйтов. Кроме того, совершенно не учитываются предполагаемые потоки данных между ПЭ

№	Маркер	Действие	Примеры
34	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.
35	Так как	Save_delete	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.*
36	Поэтому	Mdelete_save	***Поэтому*** техническая система, созданная из различных серийно изготавливаемых элементов, также будет иметь параметры и характеристики, носящие неопределенный характер и изменяющиеся в пределах некоторых интервалов.
37	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 $T(\omega) = = T(x, y, z, \omega)$. ***При этом*** *температура в каждой точке технической системы будет изменяться в некотором интервале и иметь распределение вероятностей, вообще говоря, отличное от равномерного.*
38	Для его внедрения /использования	Save	Для его внедрения в единое синтезированное трехмерное окружение создан метод рир-проекции, базирующийся на методе 3D-кеинга.

ПРИЛОЖЕНИЕ 2

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 canny, аффинные преобразования]

topic_1 | 0,08513 [Текстурный, получение, граничные точки, фон, принадлежность, выбор, ситуация, canny, 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, rothwell, 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 рас-

сматриваемом 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 неразличимо.
93:_ Использование классических метрик не выявляет преимущество ПРОГРАММной реализации МЕТОДА Canny, не пропустившей УГЛОВые точки квадрата.
97:_ Если эти функции принадлежности будут подчеркивать значимость пикселов 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 бестекстурной 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 неразличимо.

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 П4

Сравнение автоматически полученной и авторской аннотаций

Авторская аннотация	Автоматическая аннотация
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 **программ**ной 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	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**.

Окончание табл. П4

Авторская аннотация	Автоматическая аннотация
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	**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.

Примечание к табл. П4.
Жирным шрифтом выделены совпадающие фрагменты аннотаций.

ОГЛАВЛЕНИЕ

Введение 3

Глава 1. Классификация подходов 8
Глава 2. Экстрагирующие методы 18
Глава 3. Абстрагирующие методы 22
Глава 4. Гибридные методы 28
Глава 5. Scientific Text Summarizer 32
Глава 6. Униграммные тематические модели 36
Глава 7. Проблема многословных терминов 45
Глава 8. Расширенные тематические модели 52
Глава 9. Риторический анализ и преобразования графов 56
Глава 10. Сглаживание текста аннотации 63
Глава 11. Результаты экспериментов 68

Заключение 74
Список сокращений и условных обозначений 75
Список литературы 78

Приложение 1. Маркеры и коннекторы 86
Приложение 2. Примеры работы системы 99

Научное издание

Батура Татьяна Викторовна,
Бакиева Айгерим Муратовна

МЕТОДЫ И СИСТЕМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Монография

Редактор *Я. О. Козлова*
Верстка *А. С. Терешкиной*
Обложка *Е. В. Неклюдовой*

Подписано в печать 29.11.2019 г.
Формат 60 × 84/16. Уч.-изд. л. 6,9. Усл. печ. л 6,3.
Тираж 100 экз. Заказ № 316.
Издательско-полиграфический центр НГУ
630090, Новосибирск, ул. Пирогова, 2.