На правах рукописи

ОСМИНИН Павел Григорьевич

ПОСТРОЕНИЕ МОДЕЛИ РЕФЕРИРОВАНИЯ И АННОТИРОВАНИЯ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ, ОРИЕНТИРОВАННОЙ НА АВТОМАТИЧЕСКИЙ ПЕРЕВОД

Специальность 10.02.21 – Прикладная и математическая лингвистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата филологических наук

Челябинск

2016

Работа выполнена на кафедре лингвистики и межкультурной коммуникации факультета лингвистики ФГБОУ ВПО «Южно-Уральский государственный университет» (национальный исследовательский университет)

Научный руководитель: доктор филологических наук, доцент
Шереметьева Светлана Олеговна
ФГБОУ ВПО «Южно-Уральский государственный университет» (национальный исследовательский университет)

Официальные оппоненты: доктор филологических наук, доцент, заведующий кафедрой иностранных языков ФГБОУ ВО «Смоленский государственный университет»
Андреев Вадим Сергеевич

кандидат филологических наук, доцент, доцент кафедры математической лингвистики ФГБОУ ВПО «Санкт-Петербургский государственный университет»
Захаров Виктор Павлович

Ведущая организация: ФГБУН Институт проблем передачи информации им. А.А. Харкевича РАН

Защита состоится 2 ноября 2016 года в 10.00 на заседании диссертационного совета Д 212.274.15 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук в ФГБОУ ВО «Тюменский государственный университет» по адресу: 625003, г. Тюмень, ул. Республики, 9, ауд. 211.

С диссертацией можно ознакомиться в Информационно-библиотечном центре ФГБОУ ВО «Тюменский государственный университет» по адресу: 625003, г. Тюмень, ул. Семакова, 18, а также на официальном сайте ТюмГУ, код доступа: http://d21227415.utmn.ru

Автореферат разослан «____» _____ 201__ г.

Ученый секретарь
диссертационного совета
кандидат филологических наук, доцент Т.В. Сотникова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Реферируемая диссертационная работа посвящена проблеме моделирования научно-технических рефератов и аннотаций как основе автоматизации этого процесса, при котором полученный текст реферата (аннотации) не только корректно отражает содержание документа, но и обладает структурой, позволяющей избежать существенного количества ошибок при автоматическом переводе. Проблема рассматривается на примере реферирования и аннотирования статей по математическому моделированию, представляющих собой разновидность научно-технических текстов.

Актуальность темы исследования обусловлена тем, что разработка формальных моделей реферирования и аннотирования является основой автоматизации этих процессов, что настоятельно требуется необходимостью оперативной обработки все возрастающих потоков информации. Рефераты и аннотации научных статей служат приоритетным средством обмена информацией в процессе профессиональной коммуникации. При этом при постоянно возрастающих потоках информации возникает угроза обесценивания информации из-за трудностей поиска необходимых сведений среди множества доступных текстов. Несмотря на то, что исследования в области моделирования процессов реферирования и аннотирования текстов продолжаются уже более 65 лет, проблема формального получения высококачественных рефератов до сих пор не решена, что обусловлено сложностью этой задачи.

В настоящее время, когда темпы глобализации требуют все более оперативного обмена научно-технической информацией в международном масштабе, а ценность научных публикаций часто определяется их включением в престижные журналы и международные базы цитирования (например, Scopus, Web of Science), особенно остро встает проблема повышения качества автоматического перевода рефератов и аннотаций на английский язык. При всем несовершенстве автоматического перевода (АП), который, тем не менее, превращается в социальную и экономическую необходимость, качество продукции систем АП можно существенно повысить, если выявить, систематизировать и устранить из текста лингвистические явления, вызывающие ошибки при АП. Этой проблеме до сих пор не уделялось должного внимания, особенно в области моделирования рефератов и аннотаций.

Модель реферирования научно-технических текстов разрабатывается на примере научных статей в области математического моделирования до сих пор остающихся за рамками внимания исследователей-разработчиков систем реферирования. Между тем реферирование статей по математическому моделированию особенно актуально, поскольку математическое моделирование используется практически во всех отраслях науки и техники.

Степень разработанности проблемы. Исследования по моделированию и автоматизации реферирования и аннотирования возникли во второй половине XX века. Этой проблемой занималось большое количество исследователей как в нашей стране (Р.Г. Пиотровский, В.П. Леонов, Д.Г. Лахути, Э.Ф. Скороходько, С.М.

Приходько, В.А Яцко, Н.В. Лукашевич, О.А. Емашова, В.С. Ступин, О.В. Корхова, А.В. Анисимов, С.А.Тревгода, и др.), так и за рубежом (H.P. Luhn, D. Marcu, K. Ono, D. Radev, H. Saggion, L. Plaza, H.P. Edmundson, J. Kupiec, E. Lloret, J.J. Pollock, Pierre-Etienne Genest, U. Hahn, T. Strzalkowski и др.). С начала проведения исследований по автоматическому реферированию и аннотированию было разработано множество различных методов, которые можно разделить на две группы: экстрагирующие или извлекающие методы, основанные на извлечении из первичных документов наиболее информативных фрагментов и включении их в реферат в порядке следования в тексте, и абстрагирующие или генерирующие методы, предусматривающие создание нового текста, обобщающего первичные документы. Среди методов второй группы можно выделить чисто абстрагирующие методы, обобщающие текст первичного документа на достаточно высоком уровне и гибридные методы, которые сочетают техники экстракции и абстракции.

Рефераты, полученные в рамках экстрагирующих подходов, часто характеризуются низким качеством текста — бессвязностью, и низкой степенью сжатия — так как не выполняется обобщение информации и не происходит замены конкретных слов на более общие понятия.

Чисто абстрагирующие подходы потенциально способны обеспечить лучшее качество текста реферата и более высокую степень сжатия текста, но они чрезвычайно трудны для практической реализации и находятся на уровне исследовательских разработок. Сложность в реализации гибридных методов заключается в выборе наиболее удачного сочетания сторон абстракции и экстракции.

Несмотря на множество исследований, проблема разработки формальных моделей для автоматического реферирования и аннотирования еще не решена, так как естественный язык характеризуется неоднозначностью, неограниченностью и чрезвычайно сложно поддается формализации.

Цель исследования состоит в разработке общего алгоритма и основных компонентов формальной модели реферирования и аннотирования научно-технических текстов, ориентированной на генерацию корректного по содержанию текста реферата с синтаксической структурой, позволяющей избежать значительного числа ошибок при автоматическом переводе.

Поставленная цель достигается последовательным решением **задач:**

- изучение понятия «реферат» и «аннотация» в отечественной и зарубежной практике, выявление их функций, требований к составлению;
- исследование различных подходов к формализации и автоматизации реферирования и аннотирования научно-технических текстов;
- исследование различных подходов к извлечению ключевых слов;
- анализ параметров переводимости научно-технического текста и разработка правил контролируемого языка для повышения качества автоматического перевода рефератов научно-технических текстов;
- создание специализированного корпуса научных статей и соответствующих авторских рефератов и их анализ на основе количественных методов;

- разработка автоматизированной методики сопоставительного анализа полнотекстовых статей и их рефератов;
- определение лингво-статистических характеристик научных статей, рефератов и аннотаций и их соотношения в ходе их сопоставительного анализа с выявлением формальных индикаторов включения информации в реферат и аннотацию;
- разработка базы знаний модели;
- разработка правил извлечения релевантной для реферата информации из полнотекстовых документов и ее формального представления в виде шаблонов;
- разработка правил генерации текста реферата на основе шаблонов при соблюдении правил контролируемого языка;
- разработка алгоритмов моделирования и апробация модели реферирования и аннотирования научных текстов.

Объектом исследования являются структура и подъязык научно-технических текстов и соответствующих рефератов и аннотаций.

Предметом исследования является разработка модели на основе выделенных корреляций фрагментов научно-технических текстов и соответствующих рефератов и аннотаций, обусловленных особенностями структуры подъязыка.

Материалом исследования являются корпуса 137 текстов научных статей и соответствующих им текстов рефератов/аннотаций по математическому моделированию на русском языке, из которых 107 документов было использовано для построения базы знаний (объем корпуса статей — 203729 словоупотреблений без учета библиографических списков, объем корпуса рефератов — 4924 словоупотреблений), а 30 документов дополнительно использовались для апробации модели (объем корпуса статей — 99000 словоупотреблений без учета библиографических списков, объем корпуса рефератов — 2000 словоупотреблений). Научные статьи и соответствующие рефераты/аннотации были взяты из следующих журналов и сборников статей: «Вестник Южно-Уральского государственного университета. Серия: Математическое моделирование и программирование», «Математическое моделирование», «Вестник Томского государственного университета. Математика и механика», «Математические заметки», «Вестник Ивановского государственного энергетического университета», «Известия Челябинского научного центра УрО РАН».

Научная новизна работы состоит в том, что языковой материал впервые исследуется с применением совокупности современных лингвистических и компьютерных методов, что обеспечило новизну полученных результатов. Существенной новизной отличаются конкретная методика сопоставления полнотекстовых документов и авторских рефератов с помощью существующих программ автоматизированного перевода. Новой является достаточно глубокая база знаний модели, включающая информационно-концептуальную сеть в виде корневого дерева, фреймовые шаблоны для глубинного представления содержания реферата, стоп-лексикон, правила извлечения релевантной для реферата

информации, основанные не только на распределении и весе ключевых слов, но и на выделении семантических маркеров и наложении фреймовых шаблонов, а также правила генерации текстов реферата, включающие правила контролируемого языка реферата, позволяющие избежать существенного количества ошибок при АП на иностранный язык.

Актуальность и новизна исследования определяют его теоретическую и практическую значимость.

Теоретическая значимость исследования заключается в моделировании механизмов идентификации основного содержания научно-технического документа на основе достаточно глубокого (морфосинтаксического и семантического) анализа его лингвистической структуры, а также лингвистических механизмов порождения нового текста строгой функциональной направленности на основе формального представления содержания. Теоретическую значимость имеют способы представления знаний в виде информационно-концептуальных сетей и фреймовых шаблонов. Методика идентификации релевантного содержания реферата развивается путем введения в дополнение к распределению ключевых слов новых параметров различного семантического статуса, позволяющих определить релевантность квантов информации статьи для определенной информационной части реферата (тема, цель, метод, результат). Разработанная в процессе исследования методика сопоставительного анализа текстов одного языка с помощью существующих инструментов автоматизированного перевода и результаты сопоставительного анализа полнотекстовых научно-технических документов и авторских рефератов в области математического моделирования вносят определенный вклад в разработку таксономии подъязыков науки и техники, а также в развитие теории обработки естественного языка.

Практическая значимость исследования заключается в возможности создания на базе разработанной модели системы автоматического реферирования и аннотирования, с помощью которой решаются задачи облегчения и повышения оперативности оформления реферативных документов на родном и иностранных языках. Описанная модель допускает дальнейшее развитие и может быть экстраполирована на другие предметные области и национальные языки. Результаты исследования и конкретные результаты анализа подъязыка математического моделирования, а также разработанный контролируемый язык могут использоваться для разработки других типов систем автоматической обработки текста, например, информационно-поисковых систем и систем автоматического перевода. Отдельные положения работы могут применяться при обучении реферированию и аннотированию, чтении курсов по функциональной стилистике и прикладной лингвистике.

В работе использовались следующие **методы исследования**: метод сплошной выборки, описательный метод, метод статистического анализа, метод трансформаций, метод моделирования, метод сопоставительного анализа, метод экспертных оценок.

Теоретическую базу и методологическую основу исследования составили положения теории свертывания информации, приведенные в работе Д.И. Блюменау

(2002); теоретические положения информационного поиска G. Salton (1975), K.S. Jones (2004); теория риторической структуры W.C. Mann (1988); работы по исследованию подъязыков различных областей Z. Harris (1968), J. Lehrberger (1982), S.B. Johnson (1989), R.I. Kittredge (2003), N. Sager (1990); работы по автоматическому переводу и переводимости текстов С.О. Шереметьевой (2006), Л.Н. Беляевой (2013), Е.М. Мещеряковой (2013), P. Koehn (2009), K. Uchimoto (2005), S. O'Brien (2004), A. Hartley (2012); работы по формализации естественного языка В.А. Тузова (2001), О.В. Корховой (2001); работы по автоматическому извлечению ключевых слов С.О. Шереметьевой (2009), М. Гриневой (2009), W.D. Abilhoa (2014), C.B. Ali (2013), S. Rose (2010); работы по ручному реферированию В.И. Горьковой (1964), E.T. Cremmins (1982), R.E. Maizell (1978), J. E. Rowley (1988), F.W. Lancaster (2003); работы по автоматическому реферированию Р.Г. Пиотровского (1978, 1983), В.П. Леонова (1986), Н.В. Лукашевич (1998, 2009), В.А. Яцко (2002), P.-E. Genest (2011), E. Lloret (2013), M. Kumar (2009), M.G. Ozsoy (2011), L. Plaza (2008), H. Saggion (2002, 2009), D. Radev (1998), D. Marcu (1998, 1999).

На защиту выносятся следующие **положения**:

1. Создание модели реферирования и аннотирования, которая позволяет получить реферат (аннотацию) высокого качества и снимает многие проблемы его последующего автоматического перевода, обеспечивается сочетанием экстрагирующих и абстрагирующих методик на основе лингвистической базы знаний.
2. База знаний модели представляет собой набор формальных конструктов, содержащих информацию о структуре реферата, лексических, грамматических и семантических характеристиках составляющих реферат (аннотацию) элементов и индикаторах переводимости.
3. Алгоритм создания реферата (аннотации) включает две основных процедуры: а) извлечение релевантной для реферата информации с помощью метрики, которая основана на дистрибуции ключевых слов и количественных характеристик лексем-маркеров из базы знаний, и б) генерация текста, удовлетворяющего требованиям корректности и переводимости.
4. База знаний модели и алгоритмы построения реферата (аннотации) разрабатываются как на основе анализа подъязыка рефератов по математическому моделированию, так и по результатам сопоставительного анализа полного текста статей и их рефератов.
5. Сопоставительный анализ может проводиться с помощью доступных компьютерных инструментов, предназначенных для автоматизации перевода.

Достоверность и научная обоснованность теоретических и практических результатов исследования обеспечивается:

- формированием и анализом массивов текстов полных документов и их авторских рефератов значительного объема (302729 словоупотреблений и 6924 словоупотреблений, соответственно) с применением статистического, сопоставительного и описательного методов;
- созданием лингвистической базы знаний, основную часть которой составляют информационно-концептуальная сеть, набор фреймовых шаблонов, правила

извлечения релевантной для реферата информации и правила генерации текста с учетом разработанного контролируемого языка, стоп-лексикон, состоящий из четырех списков, различным образом сжимающих текст;

- положительными результатами тестирования разработанной модели реферирования путем сравнения сгенерированных ею рефератов с «золотым» корпусом авторских рефератов, с рефератами сгенерированными другими системами (ОРФО) и на основании экспертного суждения.

Апробация работы. Основные положения исследования обсуждались на заседаниях кафедры лингвистики и межкультурной коммуникации ФГБОУ ВПО «Южно-Уральский государственный университет» (НИУ), а также докладывались на международных, всероссийских, межрегиональных и региональных конференциях: конференция аспирантов и докторантов ЮУрГУ (Челябинск, 2012-2014 гг.), «Язык. Культура. Коммуникация» (Челябинск, 2014), «Прикладная лингвистика в науке и образовании» (Санкт-Петербург, 2014), «Диалог» (Московская область, 2014). Результаты работы применялись при выполнении государственного задания 2012054-ГЗ15 по созданию системы автоматического перевода рефератов и аннотаций с русского языка на английский. Основные положения исследования отражены в 11 печатных работах, 7 из которых опубликованы в журналах, входящих в перечень ВАК.

Диссертационная работа состоит из введения, трех глав, заключения, списка сокращений, списка терминов, списка литературы и трех приложений. Объем диссертации составляет 239 страниц. В диссертации содержится пять рисунков, двенадцать таблиц. Список литературы состоит из 227 источников (из них 168 источников — на иностранных языках).

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** указывается цель работы, обосновывается ее актуальность, теоретическая и практическая значимость, описываются задачи исследования, объект, предмет и материал исследования, используемые методы исследования, указываются положения, выносимые на защиту, и приводятся сведения об апробации результатов работы.

В первой главе диссертации «Проблемы автоматического реферирования, аннотирования и переводимости научных текстов|» анализируются понятия «реферат» и «аннотация» в российской и зарубежной практике, требования, предъявляемые к англоязычным и русскоязычным аннотациям и рефератам, описываются различные методы автоматического реферирования и аннотирования и подходы к извлечению ключевых слов, как основного этапа этого процесса, а также рассматриваются проблемы автоматической переводимости текстов.

Понятия «реферат» и «аннотация» трактуются несколько по-разному как в нашей стране, так и за рубежом.

Российский ГОСТ 7.9-95 «Реферат и аннотация. Общие требования» дает следующие определения.

- «реферат — краткое точное изложение содержания документа, включающее основные фактические сведения и выводы, без дополнительной интерпретации или критических замечаний автора реферата»;
- «аннотация — краткая характеристика документа с точки зрения его назначения, содержания, вида, формы и других особенностей».

Согласно ГОСТу реферат и аннотация должны выполнять следующие функции:

- давать возможность установить основное содержание документа, определить его релевантность и решить, следует ли обращаться к полному тексту документа;
- использоваться в информационных, в том числе автоматизированных системах для поиска документов и информации.

В англоязычной литературе понятиям «реферат» и «аннотация» достаточно близко соответствуют понятия «informative abstract» и «indicative abstract» соответственно. Согласно американскому стандарту ANSI/NISO Z39.14-1997 informative abstract обычно составляется для исследовательских документов, содержащих описание методов работы, ее результаты и т.д., indicative abstract составляется для работ, не содержащих описание методов работы или объемных документов, например, книги, годовые отчеты. Отмечается, что на практике обычно составляются рефераты, сочетающие стороны informative и indicative abstract. В настоящее время в англоязычной литературе достаточно часто используется более общее понятие — «summary», которое объединяет признаки реферата и аннотации.

Согласно требованиям российского ГОСТа 7.9-95 «Реферат и аннотация. Общие требования» аннотация должна включать в себя характеристику основной темы, проблемы объекта, цели работы и ее результаты. Средний объем аннотации составляет 500 печатных знаков. В реферат входит следующая информация:

- предмет, тема, цель работы;
- метод или методология проведения работы;
- результаты работы;
- область применения результатов;
- выводы;
- дополнительная информация.

В большинстве англоязычных руководств по составлению рефератов и аннотаций приводится следующая обобщенная структура реферата: цель работы, методы, результаты.

Русскоязычные и англоязычные рефераты и аннотации выполняют одинаковые функции, при этом границы понятий «реферат», «аннотация», «informative abstract», «indicative abstract» размыты и не всегда бывает понятно, чем они отличаются. Большинство источников как отечественных, так и зарубежных в качестве обязательных рекомендуют указывать следующие информационные части реферата: цель работы, методы, результаты. Текст реферата по этим требованиям должен быть от 100 до 250 слов.

В главе рассматриваются две основных группы методов автоматического реферирования:

- автоматическое реферирование, основанное на экстрагировании из первичных документов наиболее информативных фрагментов (предложения, абзацы) и включении их в реферат в порядке следования в тексте. Значимость фрагментов может определяться по различным критериям, например, по содержанию во фрагменте ключевых слов, по расположению фрагмента в исходном тексте (заголовки, подзаголовки и т.д.), по наличию сигнальных фраз. Такие методы называются извлекающими или экстрагирующими. Достоинства экстрагирующих методов — независимость от предметной области, сравнительная простота разработки: не требуется создание обширных баз знаний, проведение детального лингвистического анализа текста. Недостатки экстрагирующих методов — полученные рефераты часто являются бессвязными.
- автоматическое реферирование, предусматривающее создание нового текста, обобщающего первичные документы. Такие методы называются генерирующими или абстрагирующими.

Среди методов второй группы выделяются чисто абстрагирующие методы, обобщающие текст первичного документа на достаточно высоком уровне посредством генерации текста реферата на основе абстрактного представления смысла реферата, и гибридные методы, которые сочетают экстракцию и элементы абстракции.

При использовании чисто абстрагирующих подходов текст реферата строится алгоритмом, основываясь на лингвистических правилах обработки естественного языка и специфике 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. Преимущества абстрагирующих методов — получение реферата более высокого качества, чем при применении экстрагирующих методов. Недостатки данных методов — сложность их практической реализации, необходимость сбора большого количества лингвистических знаний.

С целью преодоления недостатков 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. 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.

Так как естественный язык очень сложен для автоматической обработки, то исследователи стараются ориентировать автоматическое реферирование для определенных предметных областей.

В главе отмечается, что извлечение ключевых слов является основным этапом реферирования и аннотирования. Основные подходы к решению этой проблемы делятся на чисто статистические и гибридные. В рамках указанных подходов можно выделить методы, требующие наличия корпуса текстов одной тематики, и методы, не требующие такого корпуса текстов.

Преимуществами чисто статистического подхода являются универсальность алгоритмов извлечения ключевых слов и отсутствие необходимости в трудоемких и времязатратных процедурах построения лингвистических баз знаний. Однако статистические методы часто не обеспечивают удовлетворительное качество результатов.

В гибридных методиках статистические методы обработки документов дополняются одной или несколькими лингвистическими процедурами (морфологическим, синтаксическим и семантическим анализами) и лингвистическими базами знаний различной глубины (словарями, онтологиями, грамматиками, лингвистическими правилами и т.д.).

В первой главе были также рассмотрены проблемы, возникающие при автоматическом переводе, и некоторые пути их решения. Выявлено, что переводимость текста системами автоматического перевода осложняется наличием как общих, так и специфических индикаторов переводимости — явлений, затрудняющих автоматический перевод.

Выделяется два набора индикаторов переводимости. В первый набор входят универсальные индикаторы, затрудняющие АП как таковой, например, синтаксическая неоднозначность предложения.

Во второй набор включают индикаторы, характерные для конкретного языка, типа текста и системы АП. Например, при переводе патентных текстов с английского языка на датский в системе PaTrans предложные группы, стоящие в начале предложения, вызывают проблемы при переводе на датский язык.

При использовании одной и той же системы АП качество перевода можно значительно повысить, сократив по возможности количество индикаторов переводимости, т.е., наложив определенные ограничения на лексико-грамматическую структуру текста, подлежащего переводу. Исправление ошибок исходного текста также дает увеличение качества автоматического перевода.

Во второй главе диссертации «Лингвистические критерии оптимальных рефератов научно-технических текстов, ориентированных на автоматический перевод» проводится анализ подъязыка рефератов предметной области «математическое моделирование», рассматриваются явления, затрудняющие автоматический перевод текстов, формулируются правила контролируемого языка, позволяющего избежать ошибок при автоматическом переводе, и проводится сопоставительный анализ полных статей и соответствующих им рефератов. Разработана специальная методика автоматизированного сопоставительного анализа, включающего в себя

использование программ памяти переводов. По результатам проведенного анализа выявляются формальные признаки извлечения информации из статьи для включения в реферат.

Анализ подъязыка рефератов по математическому моделированию показывает, что лексический состав реферата неоднороден: выделяются лексические единицы, репрезентирующие прежде всего ключевую терминологию, выраженную именными группами и глаголами. Глаголы в реферате чаще всего выполняют служебную функцию, описывая тему статьи, цель, методику и результаты исследования.

Отмечается комплексная природа текстов по математическому моделированию, которые обладают чертами научно-технических текстов как таковых, и более специфическими характеристиками математических текстов, включая при этом описание аппарата математического моделирования и описание нематематических объектов, подвергающихся моделированию.

Анализ подъязыка показывает, что грамматика подъязыка ограничена и существенно зависит от морфосинтаксических свойств глаголов (см. Таблицу 1).

Таблица 1.

Фрагмент частотного списка глаголов рефератов по математическому моделированию и их морфологическая репрезентация

Глагол	**Общая частота**	**Лич. наст. акт.**	**Лич. пасс.**	**Лич. прош. акт.**	**Прич. акт.**	**Прич. пасс.**	**Инфинит.**	**Дееприч.**
Рассматривать	44	24	13			7		
Получать	41	1	28			12		
Исследовать	28	14	11	1		2		
Предлагать	25	4	15			6		
Являться	24	21			3			
Давать	23	2				21		
Доказывать	19	3	14			2		
Использовать	18	6	3	1	2	3	1	2
Построить	18		12			6		
Показывать	17	2	13					2

Отметим, что в соответствии с определением, данным в словаре лингвистических терминов Жеребило Т.В., личными формами глагола в нашей работе считаются типичные по сочетаемости и словоизменительным характеристикам формы глагола, которые сочетаются с формой именительного падежа в функции подлежащего и изменяются по всем словоизменительным глагольным категориям. В соответствии с этим определением краткие причастия в функции сказуемого считаются личными формами глагола.

Подавляющее большинство глаголов функционирует в личной форме настоящего времени или в форме кратких причастий, а также в форме полных причастий в качестве определений. Таким образом, при построении текста

реферата, правила генерации нужно ориентировать на употребление сказуемых в личной форме настоящего времени или кратких причастий.

Анализ лингвистических и прагматико-функциональных требований к тексту реферата позволяет считать, что оптимальный в стилистическом отношении текст реферата должен обладать следующими характеристиками:

- строго упорядоченной макроструктурой — вначале должна описываться тема исследования, затем его цель, использованные методы, полученные результаты. Некоторые аспекты содержания реферата являются факультативными, поэтому при отсутствии какой-либо части макроструктуры в реферате порядок должен оставаться последовательным.
- ограниченной синтаксической сложностью. Предложения не должны содержать явлений, осложняющих автоматический перевод: вставных конструкций, эллипсиса, инверсии. При синтезе текста реферата следует учитывать рекомендации ГОСТа.

Далее на основании литературы и собственных экспериментов с онлайн-системами автоматического перевода Google и PROMT были разработаны правила контролируемого языка, соблюдение которых позволяет избегать ошибок при автоматическом переводе рефератов и аннотаций по математическому моделированию с русского языка на английский:

1) ограничить длину предложения 20 словами,

2) не использовать вставленных конструкций,

3) не допускать синтаксической омонимии, множественной сочинительной связи,

4) предложения должны содержать глагол в личной форме,

5) не допускать эллипсис,

6) не допускать дистантного расположения зависимых членов,

7) ставить определение-причастный оборот после определяемого слова,

8) ставить определение-прилагательное до определяемого слова,

9) не допускать разбиения составного и именного сказуемых вставленными выражениями,

10) использовать в русском предложении прямой порядок слов, характерный для английского языка.

11) перед существительными, требующими при переводе определенного артикля, ставить указательные местоимения или определения, например «этот», «наш» «указанный» и т.д.

При разработке правил контролируемого языка основное внимание уделялось устранению синтаксических явлений, которые вызывают неправильный перевод. Мы считаем, что построение корректного двуязычного терминологического словаря — это задача разработчиков систем автоматического перевода.

В главе были проанализированы описанные в литературе ручные и автоматизированные способы сопоставления статей и их рефератов. Ручной метод сопоставления текстов рефератов и текстов статей точнее автоматических, так как человек может распознать сложные трансформации исходного текста. К

недостаткам ручного метода сопоставления относятся высокая стоимость работы и длительность ее выполнения, поэтому исследователи стремятся к автоматизации этого процесса.

Описанные в литературе методы автоматизированного сопоставления используют вероятностные характеристики близости слов предложений статьи и реферата, такие исследования проведены в основном для английского и японского языков и требуют глубоких специфических лингвистических знаний. Поэтому в рамках настоящей работы был разработан собственный метод сопоставления с помощью программ памяти переводов (TM-программы) — SDL Trados Studio 2009.

Для пары «статья–авторский реферат» мы создали отдельный проект в SDL Trados Studio с отдельной памятью переводов, то есть проект содержал только два файла — файл полного текста статьи, разделенного программой на предложения по знакам пунктуации и ее авторского реферата, также автоматически разделенного на предложения. Полная статья исполняла роль базы данных (БД), а реферат исполнял роль текста для перевода с помощью данной БД. В поисках совпадений мы как бы «переводили» реферат с русского языка на русский. Мы установили процент нечетких совпадений (Fuzzy matches) до минимально возможного в 30%. Далее мы открыли в программе файл полной статьи и скопировали исходные предложения (на русском языке) в поле для переведенных предложений. Таким образом, мы заполнили память переводов, где предложения текста статьи сопоставлены самим себе.

Далее мы «перевели» предложения авторского реферата с помощью памяти переводов, заполненной статьей, и программа отобразила предложения статьи из памяти переводов, совпадающие с предложениями реферата. Если для предложения реферата из текста статьи предлагалось несколько вариантов совпадений предложений, то мы выбирали первые три варианта совпадения.

Для проверки корректности разработанной методики был проведен эксперимент, в ходе которого выяснилось, что явно некорректных результатов работы TM-программы оказалось мало, поэтому возможно строить корпус пар предложений «реферат-статья» для дальнейшего анализа автоматизировано, что значительно снижает трудоемкость анализа и позволяет обработать достаточно большой корпус.

В ходе анализа сопоставленных предложений статей и рефератов выяснилось, что большая часть предложений рефератов составлена авторами из предложений статьи и/или отредактированных фрагментов статьи с изменением порядка их следования. Среди трансформаций, применяемых авторами при перенесении информации из статьи в реферат, были выделены следующие:

- Изменение формы глагола, описывающего действие.
- Использование синонимов.
- Слияние предложений.
- Слияние нескольких предложений статьи в одно предложение в реферате.
- Замена термина статьи своим гиперонимом.

- Опущение в реферате вводных выражений, ссылок на формулы.
- Копирование текстового фрагмента статьи без изменения.

Наша модель разрабатывается по гибридной методике, сочетающей в себе экстрагирование и элементы абстрагирования. Реферат по нашей модели будет строиться путем извлечения текстовых фрагментов статьи, включающих релевантное для реферата содержание, и построения на их основе нового текста реферата.

В ходе дальнейшего анализа были определены лингвистические признаки (маркеры) по которым фрагмент статьи был включен в реферат. Кроме этого, было выявлено, что в статье информация о теме, цели, методе и результатах исследования может повторяться в различных разделах и в различной языковой репрезентации, в то время как в реферате каждый тип информации (тема, цель, метод, результат) представляется один раз. Порядок изложения в реферате может не соответствовать порядку представления соответствующих типов информации в статье.

В статье каждый из требуемых четырех типов информации (тема, цель, метод, результат), как правило, сопровождается языковыми маркерами, к числу которых относятся глаголы, используемые в авторских рефератах.

Анализ показал, что множество маркеров каждой информационной части (ИЧ – тема, цель, метод, результат) по своей семантике делится на четыре группы: объекты, отношения, атрибуты объектов, атрибуты отношений. При этом лексические, семантико-информационные и морфосинтаксические свойства маркеров находятся в характерной для каждой категории корреляции. Маркеры, обозначающие объекты, выражаются существительными и местоимениями; маркеры, обозначающие отношения между объектами — глаголами, а маркеры, обозначающие атрибуты объектов и отношений, выражаются прилагательными, местоимениями, наречиями.

В предложениях статьи маркеры могут функционировать либо в качестве самостоятельных лексем, либо быть частью более длинных лексических групп. В последнем случае маркеры в зависимости от категории являются частью именных групп, глагольных групп или групп прилагательных, содержащих термины предметной области математического моделирования.

Важным результатом сопоставительного анализа текстов рефератов и текстов полных статей по математическому моделированию является информация о совместном появлении, локализации и линейной последовательности маркеров различного типа во фрагментах статьи, включающих содержание, релевантное для определенных информационных частей реферата.

В третьей главе диссертации «Модель автоматического реферирования и аннотирования» описывается разработанная модель реферирования и аннотирования и ее компоненты — база знаний и алгоритм автоматического реферирования. В главе приводятся результаты апробации разработанной модели реферирования, проводится оценка пригодности рефератов, полученных в рамках модели, к автоматическому переводу.

Основными компонентами разработанной нами базы знаний модели реферирования являются: 1) знания для автоматического извлечения ключевых слов из русских научных статей, 2) стоп-лексикон, отсекающий нерелевантные для реферата части текста статьи, 3) информационно-концептуальная сеть маркеров, сигнализирующая о релевантной для реферата информации, 4) шаблоны для извлечения фрагментов статьи, содержащих релевантную для реферата информацию и предложение-клише, 5) правила трансформации выделенных фрагментов статьи в текст реферата.

Ключевые слова в нашей модели — это наиболее релевантные именные группы (ИГ) статьи. Именные группы — это самый частотный слой лексики, наиболее тесно связанный с содержанием текста. Релевантность — это количественная характеристика именной группы, которая позволяет сортировать извлеченные единицы в зависимости от задачи, стоящей перед исследователем путем вычисления значений определенных параметров.

Для извлечения ключевых слов из русских научных статей мы адаптировали для русского языка гибридный инструмент, разработанный С.О. Шереметьевой для извлечения ключевых слов из патентов на английском языке.

Следующим компонентом базы знаний модели реферирования является стоп-лексикон, состоящий из четырех списков — A (содержит слова, (например, «итак», «однако», «окончательно»), подлежащие удалению), B (содержит слова (например, «где», «если»), при обнаружении которых удаляется часть предложения от этого слова до конца предложения), C (содержит слова (например, «положим», «пусть», «обозначим»), при обнаружении которых удаляется предложение, их содержащее), D, который используется для удаления нерелевантной для реферата информации. Списки A, B, C применяются к тексту статьи на этапе предварительной обработки, а список D применяется на этапе генерации предложений из заполненных шаблонов.

Информационно-концептуальная сеть предназначена для выделения в тексте статьи маркеров, которые сигнализируют о релевантности содержащего их фрагмента статьи для реферата. Набор маркеров определен по результатам анализа подъязыка.

Информационно-концептуальная сеть представляет собой корневое дерево (Рисунок 1), которое состоит из терминальных и нетерминальных узлов и дуг, реализующих отношение включения. Корнем дерева является концепт «Реферат», в нетерминальных узлах находятся концепты, соответствующие информационным частям реферата «Тема (T)», «Цель (A)», «Метод (M)», «Результат (R)» и семантическим типам маркеров «Объект (O)», «Отношение (P)», «Атрибут Объекта (AO)» и «Атрибут Отношения (AP)».

Шаблоны — это фреймовые структуры, кодирующие знания о допустимой совместной встречаемости и линейной последовательности маркеров и других слов в релевантном для реферата фрагменте статьи, которые служат для извлечения релевантных для реферата фрагментов статьи.

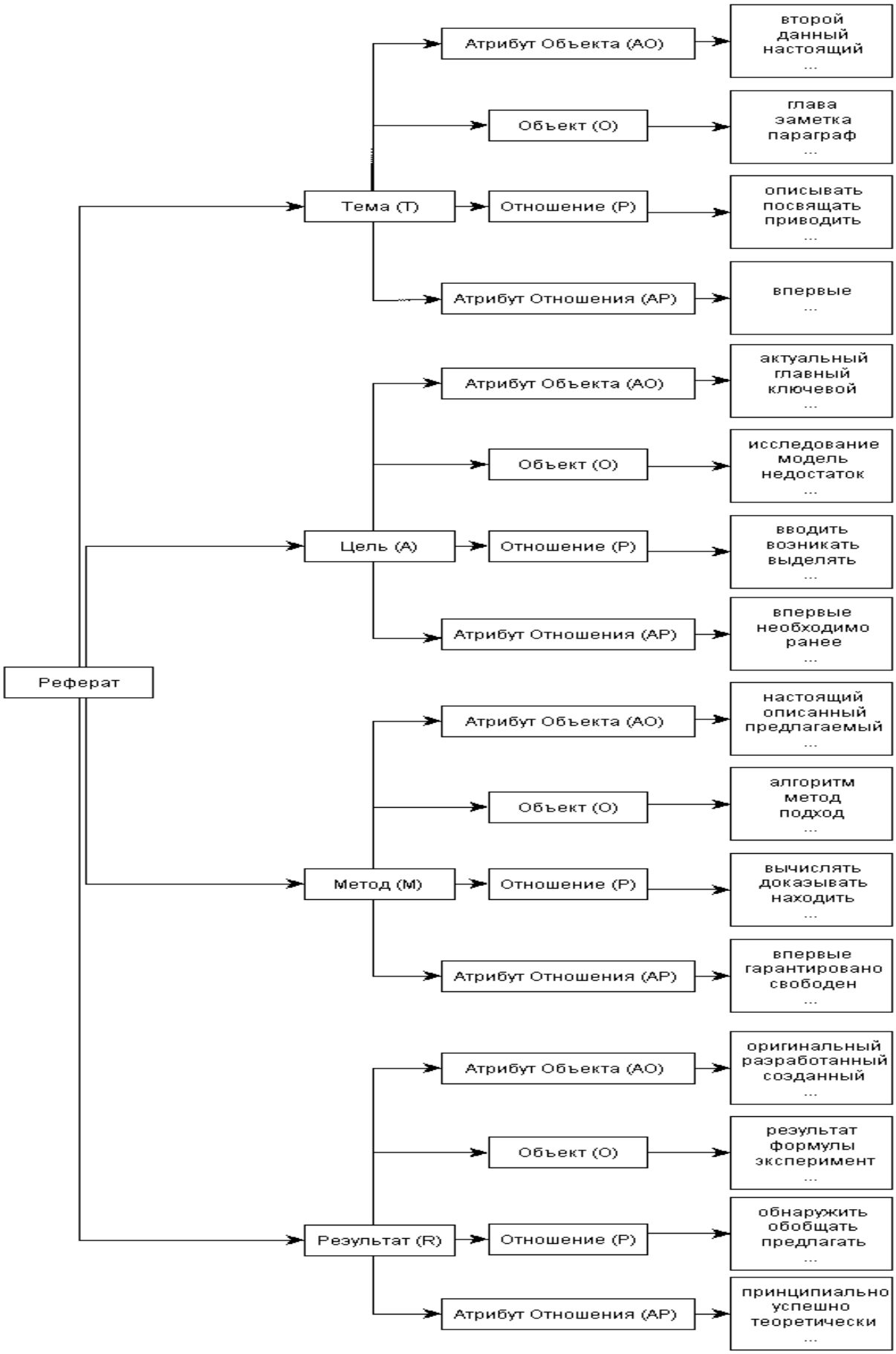

Рисунок 1 – Информационно-концептуальная сеть

Шаблоны имеют вид:

Номер шаблона	::= (натуральное число)
Шаблон	::= (ИЧ (структура))
ИЧ	::= {тема, цель, метод, результат}
Структура	::= (X Группа X … Группа …X)
X	::= {слово}
Группа	::= {NP(маркер T), VP(маркер T), AP(маркер T)}
Маркер	::= (слово-маркер_сетевой код))
Номер предложения	::= (натуральное число)
Вес	::= (натуральное число)

где

ИЧ — информационная часть реферата,

структура — структура фрагмента текста статьи,

X — цепочка из последовательных слов фрагмента, может быть пустой,

маркер — терминальный узел сети,

код — сетевой код,

T — термин,

NP — именная группа,

VP — глагольная группа,

AP — группа прилагательного,

номер шаблона — порядковый номер шаблона,

номер предложения — порядковый номер предложения, использованного для заполнения шаблона,

вес — вес шаблона, рассчитывается при взвешивании шаблона.

Предложение-клише «Рассматривается вопрос о {КС}» используется в случае, когда автоматически извлеченные ключевые слова отсутствуют в тексте сгенерированного реферата.

Пятый компонент базы знаний — правила трансформации выделенных фрагментов статьи в текст реферата и правила контролируемого языка.

Правила трансформации учитывают значения следующих параметров: принадлежность шаблона к ИЧ, вес шаблона, наличие в шаблоне маркеров определенного семантического статуса, морфологической формы глагола, лексем, обозначающих элементы статьи (например, параграф, раздел, таблицы).

Правила контролируемого языка служат для адаптации текста реферата для повышения качества его АП.

Алгоритм автоматического реферирования/аннотирования состоит из 4 основных процедур (блок-схема алгоритма приведена на Рисунке 2):

- предварительной обработки текста статьи,
- анализа текста статьи,
- отбора фрагментов статьи, содержащих релевантную для реферата информацию,
- генерации текста реферата.

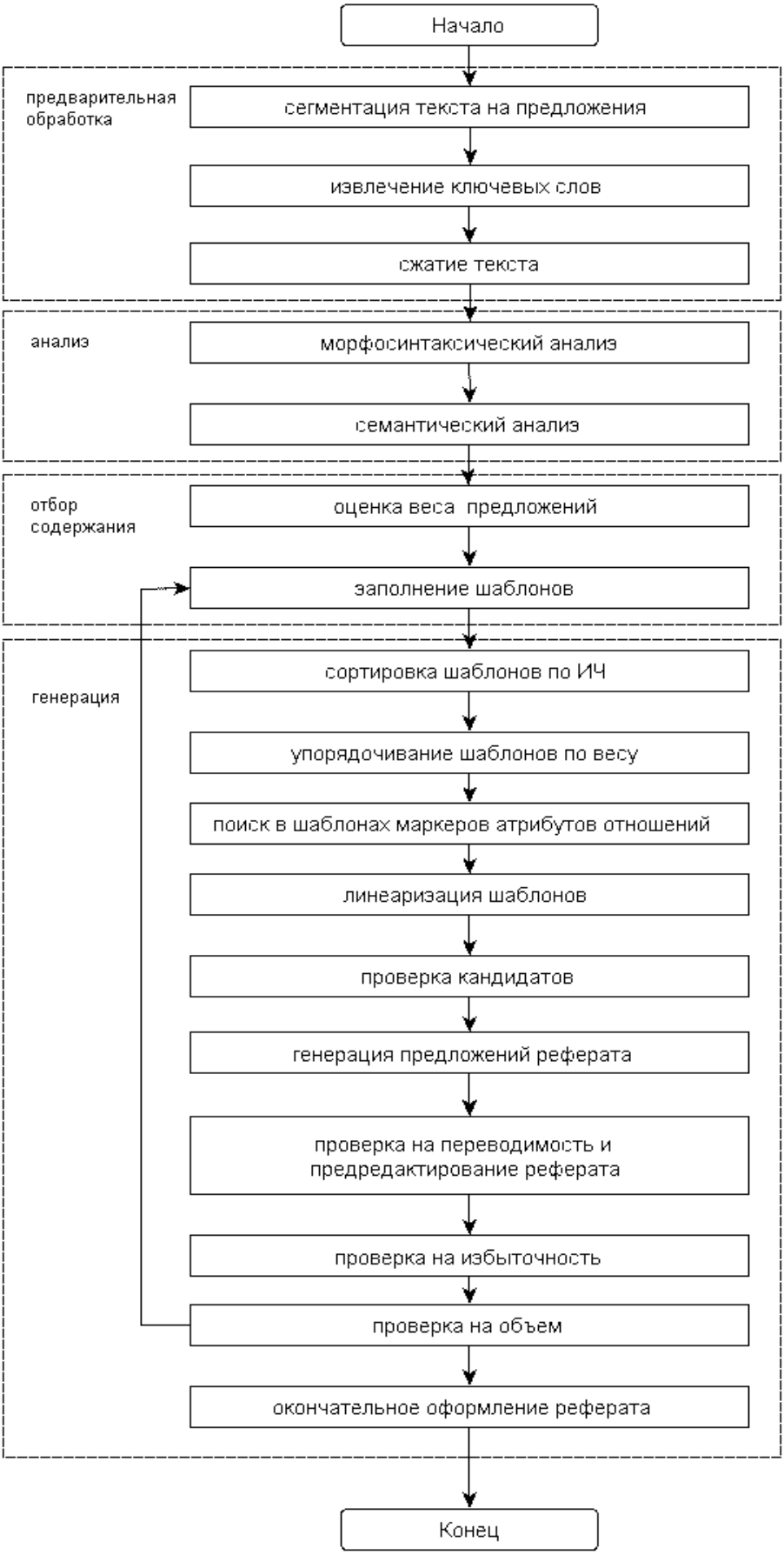

Рисунок 2 – Алгоритм автоматического реферирования

Процедура предварительной обработки состоит из трех последовательных подпроцедур — сегментации текста на предложения, извлечения ключевых слов, сжатия текста.

Первой выполняется подпроцедура сегментации текста полной статьи на предложения. Для сегментации текста на предложения могут использоваться TM-программы, например, SDL Trados Studio 2009.

Далее с помощью описанного ранее экстрактора из текста извлекаются десять наиболее релевантных ключевых слов, и запоминается их вес (релевантность).

Затем выполняется подпроцедура сжатия текста, которая удаляет из текста неинформативные для реферата слова и фрагменты статьи путем последовательного применения стоп-листов A, B, C. Кроме обработки стоп-листами на этом шаге удаляются математические формулы и предложения, содержащие менее пяти слов (словом считается текстовый фрагмент от пробела до пробела).

На этом этапе предварительная обработка заканчивается, после чего выполняется анализ оставшихся фрагментов статьи.

Процедура анализа состоит их двух подпроцедур – морфосинтаксического анализа и семантического анализа.

На этапе морфосинтаксического анализа происходит выделение лексических групп и определение их принадлежности к частям речи.

Именные группы размечаются на основе динамических знаний, полученных с помощью экстрактора, использованного для извлечения ключевых слов, поскольку этот экстрактор извлекает из текстов любой длины все именные группы, а не только ключевые. Список извлеченных именных групп накладывается на текст статьи, совпадающим фрагментам присваивается метка именной группы. Именным группам, которые являются ключевыми, приписываются метки именной группы, ключевого слова и его вес (релевантность).

Завершить морфосинтаксический анализ неразмеченных фрагментов можно с помощью доступного онлайн программного обеспечения, например, aot.ru или инструментов Russian tagging resources.

Далее на размеченном таким образом тексте статьи выполняется семантический анализ с помощью информационно-концептуальной сети. Слова текста сравниваются с терминальными узлам сети, и в случае совпадения этим словам присваивается метка маркера и сетевой код — путь от терминального узла-маркера к вершине сети.

Третья процедура алгоритма реферирования — процедура отбора содержания состоит из подпроцедур оценки веса (релевантности) предложений и заполнения шаблонов.

Идентификация фрагментов статьи с возможно релевантным для реферата содержанием осуществляется путем

а) идентификации маркеров с помощью сопоставления лексического состава статьи со множеством терминальных узлов информационно-семантической сети,

б) определением фрагментов статьи, соответствующих шаблонам базы знаний,

в) оценки веса фрагмента на основе экспериментально разработанной метрики.

Вес предложения (релевантность) определяется по следующей формуле:

$$W = 5N + M_i + K_i \quad (1)$$

где

W — вес предложения (релевантность)

N — количество ключевых слов в предложении, множитель 5 был получен эмпирически

M_i — вес всех маркеров в предложении (вес одного маркера равен 10)

K_i — вес всех ключевых слов в предложении (вес определен экстрактором).

Мы определили, что для малых текстов (до 9000 знаков с пробелами) порог отбора составляет 10 наиболее релевантных предложений, для больших текстов — 10% наиболее релевантных предложений.

Подпроцедура заполнения шаблонов осуществляется сопоставлением шаблонов и отобранных наиболее релевантных предложений. Предложения просматриваются слева направо. Неоднозначность маркеров разрешается при заполнении определенных слотов шаблона, то есть у маркера остается тот код, который присутствует в слоте шаблона. При заполнении слота X разрешение однозначности маркеров не требуется, поскольку весь фрагмент X переносится в шаблон целиком.

Процедура генерации текста состоит из трансформации заполненных шаблонов в предложения реферата, в соответствии с правилами генерации предложений, по которым фрагменты текста, заполняющие слоты шаблона выстраиваются в линейном порядке и далее проверяются на наличие индикаторов переводимости и, в случае необходимости, редактируются. На этой же стадии выполняется проверка на избыточность текста реферата и его объем.

После проверки предложения на индикаторы переводимости происходит проверка текста реферата на избыточность — присутствие предложений, выражающих одинаковое содержание. Предложения реферата сравниваются на сходство при помощи ТМ-программы следующим образом: в программу загружается текст реферата, предложения сопоставляются самим себе (таким образом, у каждого предложения есть 100% совпадение), и просматриваются на нечеткие совпадения. Если для предложения есть нечеткое совпадение, то такое предложение считается избыточным и не включается в окончательный текст реферата.

Затем происходит проверка объема реферата. Если объем реферата меньше рекомендуемого ГОСТом, то процедура рекурсивно возвращается на этап наложения шаблонов на взвешенные предложения, пока объем реферата не достигнет рекомендуемого объема. Шаблоны накладываются на предложения, которые не вошли в порог отбора на этапе взвешивания.

На заключительном шаге перед текстом реферата ставится авторский заголовок статьи, и после текста реферата добавляется блок ключевых слов, состоящий из пяти наиболее релевантных, поскольку по рекомендациям журналов требуется от трех до пяти ключевых слов. При этом более короткие ключевые слова,

являющиеся частью более длинных ключевых терминов, не повторяются. Если какие-либо выделенные ключевые слова отсутствуют в тексте сгенерированного реферата, то они заполняют слот {КС} предложения-клише ИЧ «Тема» «Рассматривается вопрос о {КС}».

В главе рассматриваются две группы методов по оценке рефератов — внутренние и внешние.

Внутренние способы оценки направлены на оценку качества самого реферата, например, учитываются связность текста, его удобочитаемость, грамматическая правильность, информативность реферата.

Внутренняя оценка часто может производиться как с привлечением экспертов — профессионалов в конкретной области науки и техники, так и самими разработчиками систем автоматического реферирования.

В случае, когда качество реферата оценивается разработчиками систем, автоматически полученный текст реферата сравнивается с «золотым стандартом» — рефератами составленными экспертами.

Еще один способ внутренней оценки качества реферата заключается в сравнении работы конкретной системы автоматического реферирования с результатами работы других систем реферирования, которое также может осуществляться вручную или автоматически.

Внешняя оценка качества систем автоматического реферирования заключается в том, насколько автоматически полученный реферат пригоден для решения других проблем, например проблемы понимания полного текста документа по его реферату, классификации документа по его реферату и т. д.

В нашем исследовании оценка разработанной модели реферирования проводилась на основе экспертной оценки:

а) путем сравнения полученных по разработанной модели рефератов с «золотыми» авторскими рефератами и

б) путем сравнения полученных по разработанной модели рефератов с рефератами, автоматически сгенерированными системой ОРФО для одних и тех же статей.

Экспертная оценка результатов реферирования показала, что рефераты, сгенерированные в рамках нашей модели, в большей степени отвечают требованиям ГОСТа и ближе к «золотому» авторскому реферату, чем рефераты системы ОРФО, что было подтверждено экспертами в области математического моделирования.

Оценка пригодности текстов рефератов для автоматического перевода проверялась с помощью системы автоматического перевода PROMT, без учета корректности перевода терминологии. Результаты АП рефератов, сгенерированных по нашей модели, показали достаточно высокое качество перевода.

В **Заключении** приводятся общие итоги и выводы по проведенному исследованию.

Итогом диссертационного исследования можно считать моделирование механизмов идентификации основного содержания научно-технического документа на основе достаточно глубокого (морфосинтаксического и семантического) анализа его лингвистической структуры, а также

лингвистических механизмов порождения нового текста строгой функциональной направленности на основе формального представления содержания. В ходе данного исследования была разработана методика сопоставительного анализа текстов одного языка с помощью существующих инструментов автоматического перевода.

В ходе исследования был произведен анализ подъязыка рефератов и статей по математическому моделированию и сопоставительный анализ рефератов и полных статей. На основе результатов анализа были разработаны база знаний и алгоритм автоматического реферирования и аннотирования.

На базе разработанной модели возможно создание системы автоматического реферирования и аннотирования. Результаты исследования и конкретные результаты анализа подъязыка математического моделирования, а также разработанный контролируемый язык могут использоваться для разработки других типов систем автоматической обработки текста, например, информационно-поисковых систем и систем автоматического перевода. Отдельные положения работы могут применяться при обучении реферированию и аннотированию, чтении курсов по функциональной стилистике и прикладной лингвистике.

Предложенную модель реферирования и аннотирования возможно улучшить. Для увеличения степени абстракции возможно составить небольшой тезаурус понятий предметной области «математическое моделирование» и включить его в модель. Далее необходимо экспериментально проверить возможность работы модели в других предметных областях. Указанные предложения станут направлением дальнейшей работы.

Основные положения диссертации отражены в следующих публикациях.

В изданиях, рекомендованных ВАК РФ:

Осминин, П.Г. Современные подходы к автоматическому реферированию и аннотированию / П.Г. Осминин // Вестник ЮУрГУ. Серия: Лингвистика. – 2012. – № 25. – С. 134-135.

Шереметьева, С. О. База знаний для автоматического определения содержания реферата / С.О. Шереметьева, П.Г. Осминин // Вестник ЮУрГУ. Серия: Лингвистика. – 2013. – Т.10. – №2. – С. 77-81.

Бабина, О.И. Память переводов в обучении переводчиков / О.И. Бабина, П.Г. Осминин // Вестник ЮУрГУ. Серия: Образование. Педагогические науки. – 2013. – Т.5. – №3. – С. 98-108.

Шереметьева, С. О. К вопросу об электронных ресурсах профессиональной лексики / С.О. Шереметьева, П.Г. Осминин, Е.С. Щербаков // Вестник ЮУрГУ. Серия: Лингвистика. – 2014. – Т.11. – №1. – С. 57-63.

Осминин, П.Г. О возможности использования CAT-программ для автоматизации одноязычных лингвистических исследований / П.Г. Осминин // Вестник Орловского государственного университета. Серия: Новые гуманитарные исследования. – 2014. – №2(37). – С. 230-232.

Осминин, П.Г. Модель автоматического реферирования на основе базы знаний, ориентированная на автоматический перевод / П.Г. Осминин // Вестник ЮУрГУ. Серия: Лингвистика. – 2014. – Т.11. – №2. – С. 65-69.

Шереметьева, С.О. Методы и модели автоматического извлечения ключевых слов / С.О. Шереметьева, П.Г. Осминин // Вестник ЮУрГУ. Серия «Лингвистика». – 2015 – Т. 12. – № 1. – С. 76–81.

Прочие публикации по теме диссертации:

Осминин, П.Г. Анализ специальных текстов, ориентированный на построение системы автоматизированного реферирования / П.Г. Осминин // Наука ЮУрГУ: материалы 65-й научной конференции. Секции социально-гуманитарных наук: в 2 т. — Челябинск : Издательский центр ЮУрГУ, 2013. —Т. 1. — С. 260–263.

Osminin, P.G. A summarization model based on the combination of extraction and abstraction / P.G. Osminin // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, 2014. — С. 461–471.

Осминин, П.Г. Создание шаблонов для извлечения информации при автоматическом реферировании текста / П.Г. Осминин // Прикладная лингвистика в науке и образовании. Сборник трудов VII международной научной конференции. 10–12 апреля 2014. г., — СПб. — С. 157-161.

Осминин, П.Г. Анализ структуры научного текста и показатели его переводимости / П.Г. Осминин // Язык. Культура. Коммуникации. – 2014. – №1. – Режим доступа: http://journals.susu.ru/lcc/article/view/20/111