На правах рукописи

Тревгода Сергей Александрович

МЕТОДЫ И АЛГОРИТМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТА НА ОСНОВЕ АНАЛИЗА ФУНКЦИОНАЛЬНЫХ ОТНОШЕНИЙ

Специальность: 05.13.01 Системный анализ, управление и обработка информации (технические системы)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург - 2009

Работа выполнена в Санкт-Петербургском государственном электротехническом университете "ЛЭТИ" им. В.И. Ульянова (Ленина)

Научный руководитель –
кандидат технических наук, доцент Сабинин Олег Юрьевич

Официальные оппоненты:
доктор технических наук, профессор Фетисов Владимир Андреевич
кандидат технических наук, доцент Власенко Сергей Владимирович

Ведущая организация – Открытое акционерное общество «Научно-технический комплекс «Ленэлектронмаш»

Защита состоится "___" _____ 2009 г. в ____ часов на заседании совета по защите докторских и кандидатских диссертаций Д 212.238.07 Санкт-Петербургского государственного электротехнического университета "ЛЭТИ" им. В.И. Ульянова (Ленина) по адресу: 197376, Санкт-Петербург, ул. Проф. Попова, 5

С диссертацией можно ознакомиться в библиотеке университета

Автореферат разослан " ___" _____ 2009 г.

Ученый секретарь
совета по защите докторских
и кандидатских диссертаций
Д 212.238.07 Цехановский В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Электронная информация играет все большую роль во всех сферах жизни современного общества. В последние годы объем научно-технической текстовой информации в электронном виде возрос настолько, что возникает угроза обесценивания этой информации в связи с трудностями поиска необходимых сведений среди множества доступных текстов. Развитие информационных ресурсов Интернет многократно усугубило проблему информационной перегрузки. В этой ситуации особенно актуальными становятся методы автоматизации реферирования текстовой информации, то есть методы получения сжатого представления текстовых документов – рефератов (аннотаций).

Постановка проблемы автоматического реферирования текста и соответственно попытки ее решения с использованием различных подходов предпринимались многими исследователями. История применения вычислительной техники для реферирования насчитывает уже более 50 лет и связана с именами таких исследователей, как Г.П. Лун, В.Е. Берзон, И.П. Севбо, Э.Ф. Скороходько, Д.Г. Лахути, Р.Г. Пиотровский и др. За эти годы выработаны многочисленные подходы к решению данной проблемы, которые достаточно четко подразделяются на два направления:

- автоматическое реферирование, основанное на экстрагировании из первичных документов с помощью определенных формальных признаков «наиболее информативных» фраз (фрагментов), совокупность которых образует некоторый экстракт;
- автоматическое реферирование, основанное на выделении из текстов с помощью специальных информационных языков наиболее существенной информации и порождении новых текстов (рефератов), содержательно обобщающих первичные документы.

В России исследования в области автоматического реферирования в настоящее время, главным образом, ведутся в рамках первого направления с использованием статистических методов, смысл которых заключается в отборе предложений с наибольшим весом, который рассчитывается на основе частоты появления слова в тексте или месторасположения предложения, для включения их в реферат. В настоящее время известны только две системы, позволяющие получать аннотации на русском языке: *TextAnalyst* и встроенная функция в пакете *Microsoft Office – Autosummarize*. Обе эти системы относятся к классу систем, использующих различные варианты статистических методов. Согласно исследованиям в области компьютерной лингвистики текст, по своей природе, нелинеен, и его структура определяется особенностями внутренней организации единиц текста и закономерностями взаимосвязи этих единиц в рамках текста как цельного сообщения. Как показала практика, различные статистические методы недостаточно эффективны, так как они интерпретируют текст в виде набора линейно упорядоченных слов, словосочетаний и предложений, игнорируя при этом лингвистическую взаимосвязанность естественного языка, что приводит к потере значимой информации.

Исследования в области автоматической обработки текстов в Европе и США привлекают внимание крупнейших частных фирм и государственных организаций

самого высокого уровня. Существует большое количество систем, разработанных, в основном, специалистами университетских центров и используемых ими для своих нужд. В этих системах предлагаются нетрадиционные решения (отличные от статистических методов), основанные на построении лексических цепочек, концептуальных графов, а также эффективных формализмов описания структуры текста. Однако все эти методы ориентированы на учет особенностей конкретных языков, в основном, английского языка, и не могут быть непосредственно применены для автоматического реферирования текстов на русском языке. Кроме того, большинство разработок носят коммерческий характер, в связи с чем принцип их работы авторами не раскрывается.

Таким образом, актуальным является создание новых эффективных методов и алгоритмов, учитывающих нелинейную и иерархическую природу текста и позволяющих получать сжатое представление текстовых документов на русском языке.

Целью диссертации является разработка новых эффективных методов и алгоритмов, учитывающих нелинейную и иерархическую природу текста, для автоматизации реферирования научно-технических текстов на русском языке.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

- Провести анализ современных подходов и методов, применяющихся при решении задачи автоматического реферирования текста.
- Разработать метод формализованного описания структуры научно-технического текста на русском языке, позволяющий автоматизировать процесс реферирования.
- Разработать алгоритм определения функциональных отношений между фрагментами текста.
- Разработать алгоритм построения структуры на основе множества функциональных отношений между фрагментами текста.
- Реализовать систему автоматического реферирования текста на основе разработанных алгоритмов и провести оценку эффективности разработанных методов и алгоритмов.

Методы исследования. Теоретической и методологической основой работы послужили: теория риторической структуры (ТРС), теория предикатов, компьютерная лингвистика, метод экспертных оценок и современные технологии программирования.

Научные положения, выносимые на защиту.

- Метод формализованного описания структуры научно-технического текста на русском языке.
- Алгоритм определения функциональных отношений между фрагментами текста на основе анализа ключевых фраз.
- Алгоритм построения структуры текста на основе множества функциональных отношений между фрагментами текста**.**

Научная новизна.

- Метод формализованного описания структуры текста, основанный на исполь-

зовании ТРС, отличается учетом нелинейной и иерархической природы текста, что позволяет повысить качество автоматического реферирования научно-технического текста на русском языке. Метод формализованного описания включает в себя определение критерия корректности структур текста, определение характеристик структуры текста и ограничений на корректные структуры текста.

- Алгоритм определения функциональных отношений между фрагментами текста отличается использованием разработанного узкоспециализированного словаря ключевых фраз русского языка и процедурами анализа отношений внутри них, что позволяет уменьшить избыточность информационного обеспечения систем автоматического реферирования за счет отказа от использования словарей и баз знаний общего назначения.

- Алгоритм построения структуры текста на основе множества функциональных отношений между фрагментами текста отличается учетом неоднозначности отношений внутри ключевых фраз русского языка посредством генерации альтернативных множеств вариантов корректных структур текста с помощью разработанных правил вывода и выбора предпочтительной альтернативы по критерию совокупной метрики, что позволяет автоматизировать процесс получения релевантной структуры текста.

Практическая ценность работы заключается в следующем:

- разработанное алгоритмическое и программное обеспечение позволяет строить системы автоматического реферирования научно-технического текста для русского языка, учитывающие нелинейную и иерархическую природу текста, что позволяет повысить качество получаемых аннотаций;

- реализована система автоматического реферирования научно-технического текста для русского языка на основе разработанного алгоритма, не требующая избыточного информационного обеспечения за счет отказа от использования обширных словарей и баз знаний общего назначения.

Внедрение и реализация результатов. Достоверность научных положений, результатов и выводов подтверждается корректным использованием математического аппарата, результатами вычислительных экспериментов по разработанным методам, алгоритмам и программам, обсуждением полученных результатов на научных конференциях, а также результатами использования и внедрения.

Полученные научные результаты внедрены и используются в Информационно-логистическом центре при Северо-Западном заочном техническом университете, в ЗАО «Абсолют» г. Санкт-Петербург, о чём имеются соответствующие акты.

Апробация работы Основные положения диссертационной работы докладывались и обсуждались на следующих конференциях:

- XII международная конференция «Современное образование: содержание, технологии, качество, Россия, Санкт-Петербург, июнь 2006г.

- XI международная научно-практическая конференция «Системный анализ в проектировании и управлении» Россия, Санкт-Петербург, июнь 2007г.

- XII международная конференция «Системный анализ в проектировании и управлении» Россия, Санкт-Петербург, июнь 2008г.

- XI международная конференция по мягким вычислениям и измерениям (SCM'2008) Россия, Санкт-Петербург, июнь 2008г.
- 62-я международная научно-техническая конференция «Системный анализ, управление и обработка информации» Россия, Санкт-Петербург, апрель 2009г.
- XШ международная научно-практическая конференция «Системный анализ в проектировании и управлении» Россия, Санкт-Петербург, июнь 2009г.

Публикации. Основные теоретические и практические результаты диссертации опубликованы в 10 статьях и докладах, среди которых 2 публикации в изданиях, рекомендованных ВАК, одна статья в других изданиях и 7 докладов на международных научно-технических конференциях.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав с выводами и заключения, изложена на 112 страницах машинописного текста, включает 26 рисунков, 26 таблиц, 4 приложения и содержит список литературы из 115 наименований, среди которых 96 отечественных и 19 иностранных изданий.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность темы диссертации, сформулированы цели и задачи работы, раскрыты основные пункты научной и практической ценности выполняемой работы, перечислены основные положения, выносимые на защиту и приведено краткое содержание глав.

В первой главе рассматриваются основные подходы к автоматическому реферированию текста. Дается обзор существующих методов автоматического реферирования, анализируются преимущества и недостатки существующих систем. Обосновываются и конкретизируются цель и задачи исследования.

Проведенный анализ известных работ в области автоматического реферирования показал, что существует два основных подхода к аннотированию:

1) извлечение из исходного текста всех «нужных» предложений (экстракция);

2) генерация реферата на основе использования методов искусственного интеллекта (абстракция).

Выполненный анализ существующих подходов к автоматическому реферированию текстов показал, что при использовании первого подхода (экстракции) результат обработки одного или нескольких документов представляется как набор предложений. Среди этого набора система выбирает те, которые в наибольшей степени подходят под заданный критерий, то есть являются более релевантными. Результатом является подмножество предложений исходного текста. Реферирование путем абстракции использует более сложные лингвистические алгоритмы, при этом выходом является не просто набор предложений из исходного текста, а порождается новый текст (реферат), содержательно обобщающий первичные документы. В этом случае для подготовки краткого изложения информации требуются мощные вычислительные ресурсы для систем обработки естественных языков, в том числе грамматики и словари для синтаксического разбора и генерации естественно-языковых конструкций. Кроме того, для реализации этого метода нужны онтологические справочники, отражающие

соображения здравого смысла и понятия, ориентированные на предметную область, для определения наиболее важной информации.

В результате анализа материалов, посвященных современным исследованиям в области автоматического реферирования текстов, выявлено, что статистические методы анализа текста, на которых до настоящего времени были сконцентрированы усилия разработчиков систем автоматического реферирования, достигли своего естественного предела. Системы, использующие вариации статистических методов анализа, не учитывают лингвистическую взаимосвязанность и нелинейность естественного языка, что объясняется, прежде всего, отсутствием эффективных методов описания структуры текста.

Структура текста определяется особенностями внутренней организации единиц текста и закономерностями взаимосвязи этих единиц в рамках текста как цельного сообщения. Каждый текст имеет функционально-стилевую ориентацию (научный текст, художественный и др.) и обладает стилистическими качествами, диктуемыми данной ориентацией.

В данной работе предлагается подход к решению задачи автоматического реферирования научно-технического текста на русском языке на основе учета особенностей структуры текста. Научной задачей в рамках предлагаемого подхода является разработка метода описания структуры текста и алгоритма автоматического реферирования, включающего в себя алгоритм определения функциональных отношений между фрагментами текста и алгоритм построения структуры текста, позволяющих повысить качество автоматического реферирования научно-технического текста на русском языке.

Вторая глава посвящена разработке метода формализованного описания структуры текста, на базе которого будет разработан алгоритм автоматического реферирования текста.

Текст состоит из функционально значимых частей. Эти части являются элементами, из которых строятся более крупные части и целые тексты. Текст не является линейной последовательностью единиц. Напротив, текст организован иерархично: элементарные единицы объединяются в единицы большего объема, те объединяются между собой и так до уровня целого текста. Для объединения единиц любого объема существует общий, единый набор структурных связей.

Предлагаемый метод формализованного описания структуры текста учитывает нелинейность естественного языка и позволяет автоматизировать процесс автоматического реферирования текстов. Метод основан на использовании теории риторической структуры, согласно которой любой текст может быть представлен в виде дерева, узлами которого являются элементарные текстовые элементы (ЭТЭ) или группы таких элементов, находящиеся в определенных отношениях между собой. Такие связи называются *риторическими отношениями* (функциональные отношения).

Текстовый элемент, вступающий в функциональное отношение, может играть в нем различную роль. Функциональные отношения, как правило, являются асимметричными: более значимый их компонент называется *ядром* (N), менее значимый — *сателлитом* (S). Сателлит часто может быть опущен или заменен другим при сохранении смысла. В то же время, если опущено или изменено ядро, смысл текста и от-

ношение существенно меняется. Бóльшая часть отношений асимметричны и бинарны, то есть содержат ядро и сателлит.

Функциональные отношения могут выстраиваться в деревья на основе пяти структурных схем, которые показаны на рис. 1. Большинство отношений соединяется, используя схему а). Схема г) покрывает случаи, в которых ядро соединено с несколькими сателлитами различными отношениями. Схемы б), в), д) показывают мультиядровые отношения.

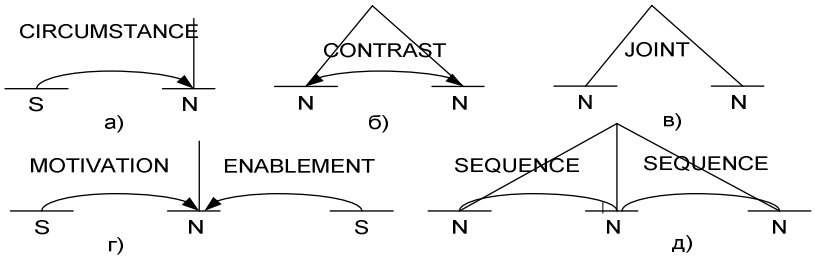

Рис. 1 Схемы функциональных отношений

При построении структуры текста, прежде всего, необходимо использовать основные положения ТРС:

- ЭТЭ представляют собой непересекающиеся части текста;
- функциональные отношения связывают текстовые элементы разного размера;
- ЭТЭ имеют в тексте различную значимость;
- структура текста может быть представлена в виде дерева.

Согласно ТРС корректными структурами текста являются такие, которые удовлетворяют следующим ограничениям:

- функциональные структуры являются деревьями, в которых элементы одного уровня представляют собой непрерывный текст;
- элементы могут быть двух типов: ядро и сателлит;
- каждый текстовый элемент может быть связан с другим только одним отношением.

С учетом этих ограничений структуры текста представляют собой деревья, смежные узлы которых представляют собой непрерывный текст.

В результате анализа основных положений ТРС установлено, что ее непосредственное применение в вычислительной модели невозможно ввиду следующих недостатков:

1) нет формальной спецификации, которая позволила бы отличить корректное дерево от некорректного;

2) нет алгоритмов для построения таких деревьев.

В связи с этим, необходимо выполнить уточнение и дополнение этой теории для описания структуры текста. Предлагаемый метод формализованного описания структуры текста включает в себя:

1. Определение критерия корректности структуры текста.
2. Определение характеристик, описывающих структуры текста.
3. Определение ограничений на корректные структуры текста.

В работе предлагается и обосновывается критерий корректности структуры текста: *если функциональное отношение лежит между двумя элементами структуры*

текста, тогда оно же лежит между, по крайней мере, двумя ключевыми составляющими этих элементов.

Стоит отметить, что ключевыми ЭТЭ являются ЭТЭ, которые играют роль ядра в функциональном отношении.

Основная идея этого критерия заключается в том, что ЭТЭ-ядра играют бóльшую роль в тексте, нежели ЭТЭ-сателлиты и, в принципе, при удалении всех сателлитов, смысл текста должен сохраниться. Если применить этот принцип рекурсивно ко всему тексту, представляя его в виде дерева, то можно получить дерево, удовлетворяющее критерию.

На следующем этапе формализации были определены характеристики структуры текста для текстового фрагмента $[l,h]$.

В работе показано, что совокупность следующих характеристик является достаточной для описания структуры текста.

- $S(l,h,status)$ показывает статус $[l,h]$. Он может иметь значения $NUCLEUS, SATELLITE$ или $NONE$.
- $T(l,h,relation\_name)$ показывает имя функционального отношения, которое лежит между прямыми потомками $[l,h]$ в дереве.
- $P(l,h,unit\_name)$ показывает имя ключевого (отражающего значимую информацию) ЭТЭ.

Статус, тип и множество ключевых узлов-потомков, которые связаны с каждым узлом, дают достаточную информацию для полного описания текстовой структуры.

Заключительным этапом формализации является определение ограничений для построения корректных структур текста. Текст представляет собой множество ЭТЭ, которые являются листьями дерева; фрагменты текста являются узлами более верхнего уровня и состоят из нескольких ЭТЭ. Допустим, имеется текст из N ЭТЭ, тогда $[l,h]$ является его фрагментом, причем l и h - левый и правый индексы ЭТЭ соответственно. В работе показано, что для генерации только корректных структур необходимо ввести следующие ограничения для текста из N ЭТЭ:

- Для каждого фрагмента $[l,h]$ предикат S имеет домен значений $NUCLEUS, SATELLITE, NONE$. Для случая, когда $l = h$, значения могут быть только $NUCLEUS, SATELLITE$:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \rightarrow \{[l = h \rightarrow (S(l,h,NUCLEUS) \vee S(l,h,SATELLITE))]$$
$$\wedge [l \ne h \rightarrow (S(l,h,NUCLEUS) \vee S(l,h,SATELLITE) \vee S(l,h,NONE))]\}$$

- Статус любого фрагмента уникален

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \rightarrow [(S(l,h,status_1) \wedge S(l,h,status_2)) \rightarrow status_1 = status_2$$

- Для каждого фрагмента $[l,h]$ предикат T имеет домен значений в виде множества функциональных отношений, соответствующих этому фрагменту

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \rightarrow \{[l = h \rightarrow T(l,h,LEAF)]$$
$$\wedge [l \ne h \rightarrow (T(l,h,NONE) \vee (T(l,h,name) \rightarrow relevant\_rel(l,h,name)))]\}$$

где $relevant\_rel(l,h,name)$ означает множество отношений, лежащих между фрагментами текста внутри $[l,h]$.

- По крайней мере, одно функциональное отношение лежит между двумя смежными фрагментами:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to [(T(l,h,name_1) \wedge T(l,h,name_2)) \to name_1 = name_2]$$

- Для каждого фрагмента $[l,h]$ предикат P имеет домен значений в виде множества ЭТЭ, из которых он состоит:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to [P(l,h,NONE) \vee P(l,h,u) \to relevant\_rel(l,h,u))]$$

- Текстовые фрагменты не пересекаются:

$$[(1 \le h_1 \le N) \wedge (1 \le l_1 \le h_1) \wedge (1 \le h_2 \le N) \wedge (1 \le l_2 \le h_2) \wedge (l_1 \prec l_2) \wedge (h_1 \prec h_2) \wedge (l_2 \le h_1)]$$
$$\to [\neg S(l_1,h_1,NONE) \to S(l_2,h_2,NONE)]$$

- Текстовый фрагмент со статусом $NONE$ не участвует в результирующем дереве:

$$[(1 \le h \le N) \wedge (1 \le l \le h)] \to [(S(l,h,NONE) \wedge P(l,h,NONE) \wedge T(l,h,NONE))$$
$$\to (\neg S(l,h,NONE) \wedge \neg P(l,h,NONE) \to \neg T(l,h,NONE))]$$

- Существует главный фрагмент, корень дерева, который покрывает весь текст:

$$(\neg S(l,N,NONE) \wedge \neg P(l,N,NONE) \to \neg T(l,N,NONE)$$

Разработанный критерий корректности структуры текста и выполненная формализация характеристик и ограничений на корректные структуры являются расширением формализации основных положений ТРС. Они определяют условия объединения фрагментов текста, позволяют минимизировать набор необходимых параметров, достаточных для полного описания структуры текста, и существенно уменьшить избыточность порождаемых альтернативных структур текста соответственно.

Третья глава посвящена разработке алгоритмов, необходимых для автоматического реферирования научно-технического текста на русском языке на основе разработанного метода.

Обобщенный алгоритм автоматического реферирования представлен на рис. 2.

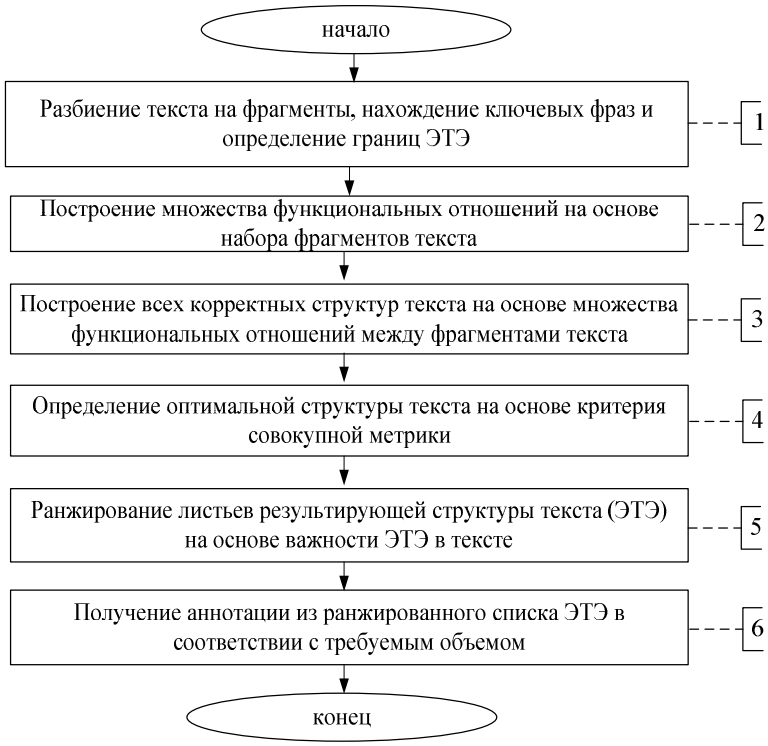

Рис. 2 Обобщенный алгоритм автоматического реферирования текста

Основными составляющими данного алгоритма являются алгоритм определения функциональных отношений между фрагментами текста на основе ключевых

фраз (блоки 1-2) и алгоритм построения структуры текста на основе множества функциональных отношений между фрагментами текста (блоки 3-4).

Первоначальной задачей при построении структуры текста является определение набора функциональных отношений между элементарными текстовыми элементами, или частями предложений. Известные подходы к решению этой задачи основаны на использовании глубокого семантического анализа текста, требующего полных баз знаний и соответствующих словарей русского языка, и до практической реализации не доведены. В данной работе на основе анализа корпуса научно-технических текстов на русском языке разработан узкоспециализированный словарь ключевых фраз русского языка, учитывающий специфику функциональных отношений между фрагментами текста, что позволяет определять множество этих отношений для научно-технических текстов на русском языке.

С помощью этого словаря алгоритм определения функциональных отношений выполняет разбиение текста на предложения, определяет границы ЭТЭ и функциональные отношения между ними. Последовательность действий при определении границ ЭТЭ представлена на рис. 3. Построение множества функциональных отношений на основе списка ЭТЭ представлено на рис. 4.

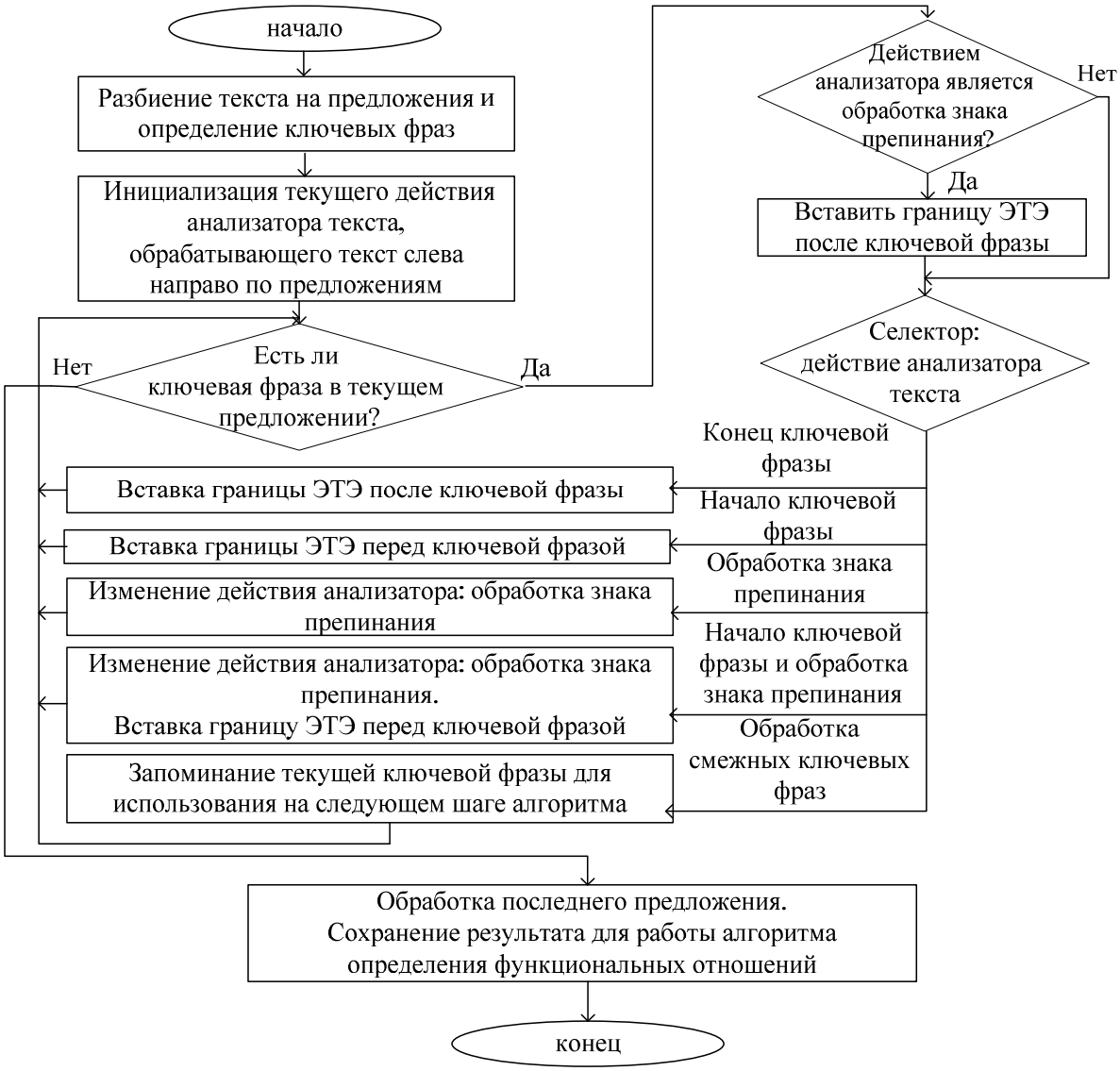

Рис. 3 Алгоритм определения границ ЭТЭ

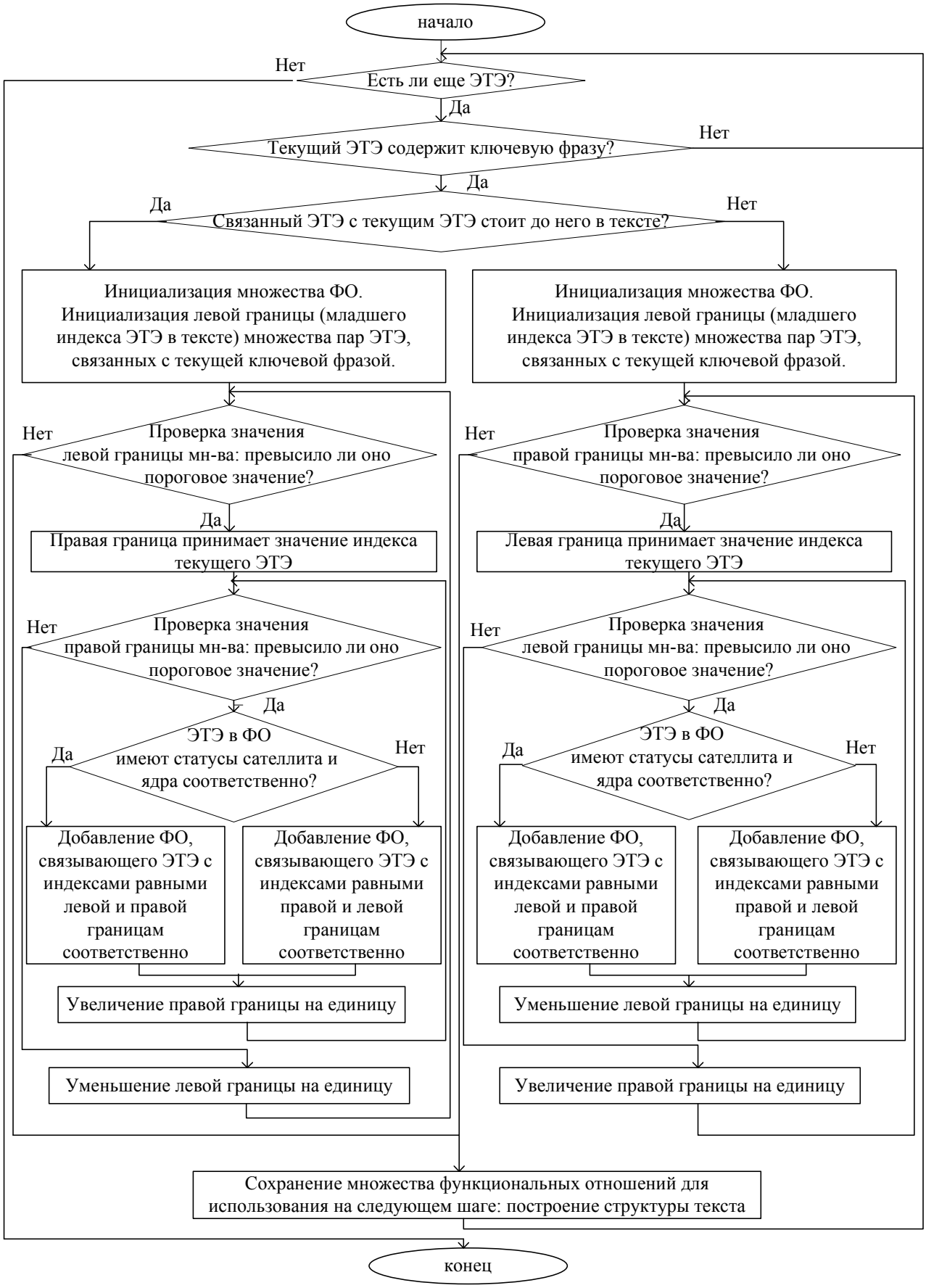

Рис. 4 Построение множества функциональных отношений на основе списка ЭТЭ

На следующем этапе был разработан алгоритм построения структуры текста, который имеет на входе сформированный набор функциональных отношений и автоматически определяет структуру текста.

Задача построения структуры текста формулируется следующим образом: дана последовательность ЭТЭ $U = u_1 u_2 ... u_n$ и множество функциональных отношений RR, которые лежат между фрагментами текста из U, найти все корректные структуры текста, исходя из линейной последовательности U.

Параметрами алгоритма построения структуры текста являются:

- множество ЭТЭ $U = u_1 u_2 ... u_n$;
- множество констант $NUCLEUS$, $SATELLITE$, $LEAF$, $NULL$;
- имена всех функциональных отношений;
- объекты типа $tree(status, type, promotion, left, right)$.

Объекты, имеющие форму $tree(status, type, promotion, left, right)$, обеспечивают функциональное представление корректных деревьев. Переменная $status$ может иметь значения $NUCLEUS$ или $SATELLITE$; $type$ содержит имя функционального отношения; $promotion$ представляет собой подмножество элементов из множества U; $left$ и $right$ могут быть либо $NULL$, либо рекурсивным определением через объект $tree$.

Идея разработанного алгоритма состоит в следующем. Изначально каждый i-й ЭТЭ ассоциирован с элементарным деревом (деревом, состоящим из одного элемента), которое имеет статус либо $NUCLEUS$ (ядро), либо $SATELLITE$ (сателлит), тип $LEAF$ (лист) и множество ключевых ЭТЭ-потомков $\{i\}$. Вначале любое отношение из множества RR может быть использовано для связи двух элементов в более сложные деревья. После построения всех элементарных деревьев структура текста формируется путем соединения смежных деревьев в бóльшие, при условии, что на каждом шаге получается корректная древовидная структура. С каждым шагом связано множество функциональных отношений, которые могут быть использованы на следующих шагах. Но как только одно из отношений было использовано, оно становится недоступным для дальнейших преобразований. Этот процесс повторяется рекурсивно до тех пор, пока не будет получена результирующая структура, покрывающая весь текст.

Реализация данного алгоритма основана на использовании системы правил вывода корректных структур текста, обоснование и формулировка которых приведены в диссертации. Данные правила вывода определяют условия объединения двух смежных фрагментов текста в более сложные структуры в различных ситуациях.

Отличительной особенностью алгоритма построения структуры текста является учет неоднозначности функциональных отношений путем генерации альтернативных корректных структур текста и выбора оптимальной структуры на основе критерия совокупной метрики в виде линейной комбинации различных индикаторов важности фрагментов текста.

Следующим этапом обобщенного алгоритма является ранжирование по важности листьев (ЭТЭ) построенного структурного дерева для всего текста (блок 5). Далее из ранжированного списка ЭТЭ выбирается их необходимое количество в соответствии с заданным объемом аннотации (блок 6) .

Разработанный алгоритм автоматического реферирования текста использует процедуру автоматического построения структуры текста на основе полученного множества функциональных отношений, что позволяет получать качественные рефераты без использования обширных словарей и баз знаний общего назначения.

В четвертой главе рассматривается система автоматического реферирования текста, построенная на основе разработанных алгоритмов, и проводится оценка эффективности разработанного метода и алгоритмов.

Процесс автоматического реферирования текста состоит из нескольких этапов, основными из которых являются следующие: анализ текста и определение функциональных отношений, построение корректных структур текста на основе этих отношений, нахождение оптимальной структуры, и затем получение аннотации.

Структура системы, реализующей разработанный алгоритм автоматического реферирования текста, представлена на рис. 5.

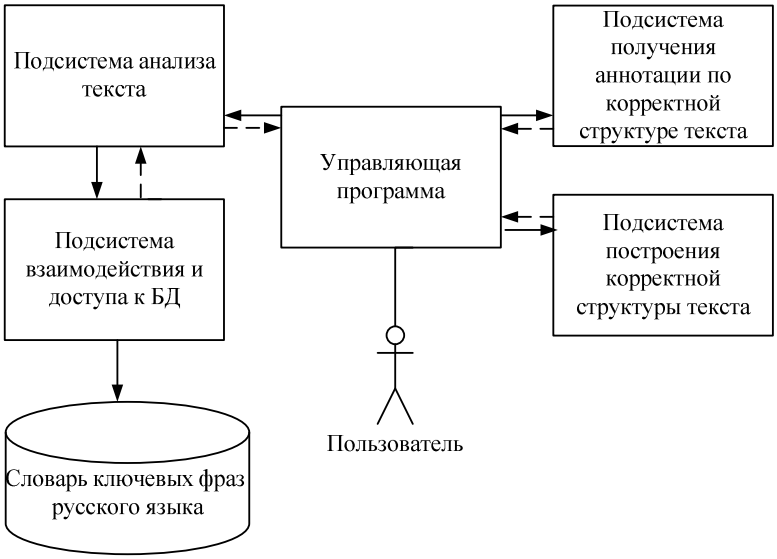

Рис 5 Структура системы автоматического реферирования текста

Реализация системы была выполнена на основе объектно-ориентированного подхода в системе программирования Java.

Эффективность разработанного метода и алгоритма автоматического реферирования оценивалась по качеству получаемых аннотаций.

Оценка качества аннотаций, получаемых с помощью разработанного алгоритма, проводилась на основе различных процедур с помощью метода экспертных оценок:

1) на основе использования эталона аннотации, составляемого группой экспертов, и формального вычисления показателей полноты и точности путем сравнения получаемых аннотаций с эталоном;

2) на основе привлечения двух групп экспертов, одна из которых составляет эталон, а другая оценивает качество аннотации по 10-балльной шкале.

При этом исследовалась зависимость качества аннотаций от объема текстов и от наличия ключевых фраз в тексте.

В соответствии с первой процедурой для оценки качества аннотаций необходимо, прежде всего, создать эталон аннотации на основе согласованного мнения боль-

шинства группы экспертов. Группа экспертов для оценки рефератов состояла из 5 научных сотрудников Информационно-логистического центра при СЗТУ. Оценка согласованности мнений экспертов рассчитывалась на основании коэффициента Кендала (коэффициент согласованности) и составила 0.7, что является достаточным для формирования эталона аннотации.

Основной задачей оценки полученной аннотации (реферата) является установление смыслового соответствия или, иными словами, семантического тождества реферата и первоисточника. Для решения данной задачи традиционно используются критерий семантической адекватности и критерий семантической эквивалентности. Первый применяется для оценки точности реферирования, второй - для оценки степени полноты отражения содержания первичного документа в реферате. Для количественной оценки критерия точности используется отношение полученных в аннотации релевантных ЭТЭ к общему количеству ЭТЭ в аннотации. Для количественной оценки критерия полноты используется отношение полученных в аннотации релевантных ЭТЭ к общему количеству релевантных ЭТЭ. Помимо этих показателей принято использовать показатель их взвешенного значения (F-параметр, или гармоническое среднее параметров полноты и точности).

В приложении к диссертационной работе приведены примеры текстов и рефератов, составленных с помощью разработанной системы, системой *TextAnalyst* и *Microsoft Autosummarize*.

Проведена сравнительная оценка разработанного метода и алгоритмов с традиционными методами, методом случайного выбора предложений для составления аннотации и усредненным выбором группы экспертов. Результаты сравнительной оценки качества методов автоматического реферирования текста на основе метода экспертных оценок по показателям полноты и точности представлены в табл. 1.

Таблица 1

Средние значения показателей качества методов автоматического реферирования

Система (метод)	*Полнота*	*Точность*	*F-параметр*
Эксперт	74.81%	80.84%	77.69%
Разработанная система (метод на основе анализа функциональных отношений)	64.81%	67.03%	66.03%
Система *TextAnalyst* (метод на основе построения семантической сети)	48.14%	44.82%	46.42%
Система *Microsoft AutoSummirize* (метод на основе подсчета статистических показателей)	35.18%	32.75%	33.92%
Метод случайного выбора	25.92%	25.92%	25.92%

Результаты экспертных оценок, при которых одна группа экспертов составляет реферат-эталон, а другая группа экспертов-оценщиков по 10-балльной системе оценивает рефераты, полученные автоматически различными системами, согласуются с оценками полноты и точности и подтверждают более высокую эффективность разработанного метода по сравнению с традиционными методами.

Исследование зависимости качества аннотаций от объема текстов и от наличия ключевых фраз в тексте показало следующее. Качество аннотаций практически не зависит от объема текста, если коэффициент встречаемости ключевых фраз в исходных текстах примерно одинаков. Качество аннотаций тем выше, чем больше коэффициент встречаемости ключевых фраз в исходном тексте. Этого следовало ожидать, ввиду того, что алгоритм основан на использовании функциональных отношений, однозначное определение которых затруднительно при отсутствии или малом количестве ключевых фраз.

Проведенные исследования показали, что качество аннотаций, полученных с помощью разработанного алгоритма, в среднем на 20% выше по сравнению с аннотациями, полученными с помощью традиционных методов для научно-технических текстов на русском языке, и при этом алгоритм имеет достаточно хорошее быстродействие, что служит основанием для его эффективного использования на практике.

ЗАКЛЮЧЕНИЕ

- Выполнен анализ современных подходов к автоматическому реферированию текстов. Установлено, что для текстов на русском языке практически реализуемыми являются методы, основанные на подходе экстракции. Анализ существующих исследований в области компьютерной лингвистики показал, что текст нелинеен по своей природе и представляет собой иерархическую структуру с определенными видами связей между фрагментами текста, несущих значимую информацию, которую можно использовать для экстрагирования.
- Предложен метод формализованного описания структуры научно-технического текста на русском языке, который отличается учетом нелинейности и иерархической природы текста, что позволяет повысить качество автоматического реферирования научно-технического текста на русском языке. Разработан критерий корректности структуры текста, выполнены формализация характеристик и ограничений на корректные структуры, которые являются расширением формализации основных положений ТРС. Они определяют условия объединения фрагментов текста, позволяют минимизировать набор необходимых параметров, достаточных для полного описания структуры текста, и существенно уменьшить избыточность порождаемых альтернативных структур текста соответственно.
- Разработан алгоритм определения функциональных отношений между фрагментами текста на основе анализа ключевых фраз русского языка, который отличается использованием разработанного узкоспециализированного словаря ключевых фраз русского языка и анализом отношений внутри них, что позволяет уменьшить избыточность информационного обеспечения систем автоматического реферирования за счет отказа от использования словарей и баз знаний общего назначения.

- Разработан алгоритм построения структуры текста на основе множества функциональных отношений между фрагментами текста, который отличается учетом неоднозначности отношений внутри ключевых фраз русского языка путем генерации альтернативных множеств вариантов корректных структур текста с помощью разработанных правил вывода и выбора предпочтительной альтернативы по критерию совокупной метрики, что позволяет автоматизировать процесс получения релевантной структуры текста.
- Проведена экспериментальная проверка предложенных метода и алгоритмов, реализованных в разработанной программной системе автоматического реферирования текста. Проведенные исследования показали, что качество аннотаций, полученных с помощью разработанного алгоритма, в среднем на 20% выше по сравнению с аннотациями, полученными с помощью традиционных методов, реализованных в системе *TextAnalyst* и встроенной функции пакета *Microsoft Office – Autosummarize.*

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Тревгода, С.А. Системы автоматического реферирования текста [Текст] / О.Ю. Сабинин, С. А. Тревгода // Приборы и системы. Управление, контроль, диагностика.- 2008. - Вып.1. - С.23-26.
2. Тревгода, С.А. Автоматизация обработки научно-технической информации [Текст] / С. А. Тревгода // Приборы и системы. Управление, контроль, диагностика. - 2009.-Вып.7. - С. 24-27.

Другие статьи и материалы конференций:

3. Тревгода, С.А. Технология автоматического реферирования технического текста [Текст] / О.Ю. Сабинин, С. А. Тревгода // Известия СПбГЭТУ «ЛЭТИ». - 2008.- № 7. - С. 25-34.
4. Тревгода, С.А. Программное обеспечение систем дистанционного управления [Текст] / С.В. Лукашевич, С. А. Тревгода // Материалы XII межд. конференции «Современное образование: содержание, технологии, качество». - Спб.: Изд-во СПбГЭТУ «ЛЭТИ», 2006. - С.25-27.
5. Тревгода, С.А. Автоматизированное проектирование функционального программного обеспечения [Текст] / С. А. Тревгода // Труды XI межд. науч.-практ. конференции «Системный анализ в проектировании и управлении». - Спб.: Изд-во Санкт-Петербургского Политех. ун-та, 2007. - С.302-304.
6. Тревгода, С.А Анализ методов автоматического реферирования технического текста. [Текст] / О.Ю. Сабинин, С. А. Тревгода // Труды XII межд. науч.-практ. конференции «Системный анализ в проектировании и управлении». - Спб.: Изд-во Санкт-Петербургского Политех. ун-та, 2008. - С.163-165.
7. Тревгода, С.А. Подход к определению множества риторических отношений для автоматического реферирования текста [Текст] / С. А. Тревгода // Труды XII межд. науч.-практ. конференции «Системный анализ в проектировании и управлении». - Спб.: Изд-во Санкт-Петербургского Политех. ун-та, 2008. - С. 166-169.

8. Тревгода, С.А Формализация процедуры построения дискурсной структуры технического текста. [Текст] / О.Ю. Сабинин, С. А. Тревгода // Материалы XI межд. конференции по мягким вычислениям и измерениям (SCM-2008). - Спб.: Изд-во СПбГЭТУ «ЛЭТИ», 2008. - С.35-38.
9. Тревгода, С.А. Алгоритм автоматического реферирования текста на русском языке [Текст] / О.Ю. Сабинин, С. А. Тревгода // Труды XIII межд. науч.-практ. конференции «Системный анализ в проектировании и управлении» - Спб.: Изд-во Санкт-Петербургского Политех. ун-та, 2009. - С. 188-190.
10. Тревгода, С.А. Методы и алгоритмы автоматического реферирования текста на основе построения и анализа дискурсных структур [Текст] / О.Ю. Сабинин, С. А. Тревгода // Труды 62-й международной науч.-техн. конференции «Системный анализ, управление и обработка информации».-Спб.: Изд-во СпбГУАП, 2009. - С. 54-57.