Московский государственный университет имени М.В. Ломоносова

На правах рукописи

Ефремова Наталья Эрнестовна

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ИЗВЛЕЧЕНИЯ ТЕРМИНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ ИЗ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2013

Работа выполнена на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова

Научный руководитель: кандидат физико-математических наук, доцент Большакова Е.И.

Официальные оппоненты: доктор технических наук, профессор Хорошевский В.Ф.

кандидат физико-математических наук Лукашевич Н.В.

Ведущая организация: Институт системного анализа РАН

Защита диссертации состоится 17 мая 2013 года в 11 часов на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМК, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за два дня по тел. (495) 939-30-10 (для оформления заявки на пропуск).

С диссертацией можно ознакомиться в Фундаментальной библиотеке МГУ имени М.В. Ломоносова. С текстом автореферата можно ознакомиться на официальном сайте факультета ВМК МГУ http://cs.msu.su/ в разделе «Наука» – «Работа диссертационного совета» – «Д 501.001.44».

Автореферат разослан 11 апреля 2013 года

Ученый секретарь
диссертационного совета В.А. Костенко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Существенная часть обрабатываемой вычислительными системами информации до сих пор представлена в виде текстов на естественном языке (ЕЯ). Число таких текстов со временем только увеличивается, в связи с чем прикладные задачи автоматической обработки текста (АОТ) не теряют своей актуальности.

Многие задачи АОТ при своем решении требуют извлечения из текста единиц, обычно – слов и словосочетаний, отражающих его содержание. Для научно-технических текстов (НТ-текстов) такими единицами являются **термины**, т.е. слова и словосочетания, называющие понятия определенной предметной области (ПО). Термины, как правило, входят в число наиболее частотных единиц НТ-текста и достаточно точно отображают его смысл.

Для автоматического извлечения терминов в настоящее время применяются методы, опирающиеся на статистические и лингвистические критерии[1]. Статистические критерии в основном используют частоты встречаемости слов и словосочетаний в обрабатываемом тексте или коллекции текстов, а также вычисляемые на основе этих частот статистические величины. Лингвистические критерии учитывают типичную синтаксическую структуру терминов и свойственные конкретной ПО конструкции, в рамках которых употребляются терминологические слова и словосочетания.

В современных системах АОТ точность распознавания терминов колеблется (в зависимости от применяемого метода) в интервале от 20% до 50%, а полнота – от 55% до 85%. При этом, основным способом повышения полноты и точности извлечения терминов является подбор нужной комбинации статистических и лингвистических критериев.

При построении компьютерных терминологических словарей и онтологий приемлемые значения полноты и точности извлечения достигаются при обработке больших коллекций текстов – в этом случае предпочтение отдается статистическим критериям. В тоже время во многих других задачах АОТ – таких, как автоматический перевод текста с одного ЕЯ на другой, реферирование и аннотирование текста, требуется анализ терминов отдельно взятого НТ-текста. Такой анализ предполагает как можно более полное

[1] Добров Б.В., Лукашевич Н.В. и др. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции". – 2003. – С. 201-210.

распознавание не только различных терминов, но и всех их вхождений в анализируемый текст. При этом возможности статистических критериев существенно ограничены, поэтому в подобных задачах необходимо ориентироваться на лингвистические критерии.

Одна из сложностей выявления различных вхождений терминов в текст связана с тем, что термины достаточно часто при употреблении видоизменяются – усекаются, сокращаются, заменяются синонимами и т.д.: *аберрация оптической системы – аберрация системы – аберрация*, *синтаксическое представление – СинП*, *вложенный файл – вложение*. Подобные **текстовые варианты** представляют собой различные формы выражения одного и того же понятия и по возможности должны быть распознаны при обработке текста. Кроме текстовых вариантов в НТ-текстах встречаются также **соединения** (комбинации) нескольких терминологических словосочетаний, которые также следует учитывать при решении прикладных задач АОТ. Типичным примером соединения терминов является фраза *естественный и искусственный отбор*, образованная из двух терминов: *естественный отбор* и *искусственный отбор*.

Большинство известных методов автоматического извлечения терминов не полностью учитывают указанные особенности употребления терминов, что существенно снижает эффективность их работы. В частности, в системах АОТ редко распознаются синонимы, текстовые варианты и соединения терминологических словосочетаний. Таким образом, проблема повышения точности и полноты автоматических методов извлечения терминов, а также их вариантов и конструкций их употребления остается до сих пор актуальной.

Цель и задачи. Основная цель настоящей диссертационной работы – повышение показателей полноты и точности автоматического извлечения из отдельно взятого НТ-текста на русском языке терминологической информации, включающей:

- общепринятые термины;
- конструкции определений новых терминов и введения их синонимов;
- текстовые варианты распознанных терминов;
- соединения нескольких терминологических словосочетаний;
- частоту употребления в тексте распознанных терминов и вариантов.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Рассмотреть современные методы извлечения терминов и существующие средства формального представления конструкций ЕЯ, исследовать их применимость для автоматического распознавания терминов, их вариантов и конструкций их употребления, типичных для русскоязычных НТ-текстов.

2. Разработать процедуры извлечения (на базе частичного синтаксического анализа) различной терминологической информации из отдельно взятого текста; предусмотреть возможность настройки процедур на новые случаи терминоупотребления.

3. Программно реализовать разработанные процедуры извлечения, и с помощью экспериментального исследования оценить качество их работы.

Поскольку объем НТ-текста может быть небольшим (научная статья, аннотация), а статистические критерии хорошо работают только для текстов значительного объема, при разработке процедур извлечения терминов и их употреблений основной упор был сделан на применение лингвистических критериев.

Методы исследования. В работе использовались методы из области искусственного интеллекта, а также информатики и программирования, в частности, методы формального представления знаний и автоматического синтаксического анализа, методики экспериментальной оценки по коллекциям текстов, а также методология объектно-ориентированного проектирования.

Научная новизна. В диссертационной работе предложен подход к разработке автоматических процедур извлечения из текста терминологической информации на базе формализации в виде лексико-синтаксических шаблонов лингвистических особенностей употребления терминов. По результатам проведенного исследования эффективности разработанных процедур предложена стратегия объединения результатов их работы, позволяющая улучшить показатели точности и полноты извлечения терминов из отдельно взятого НТ-текста, и в том числе – получать более точную информацию о частоте их употреблений в тексте.

Практическая значимость. Предложенный в диссертации подход к извлечению терминологической информации из НТ-текста и разработанные в его рамках процедуры и стратегия извлечения могут быть использованы при решении прикладных задач АОТ, в которых требуется по возможности точное и полное распознавание различных употреблений терминов в тексте. К таким задачам относятся реферирование и аннотирование НТ-текстов, построение

глоссариев и предметных указателей документа, создание и обновление машинных терминологических словарей и тезаурусов.

Применение в разработанных процедурах в качестве входных данных наборов лексико-синтаксических шаблонов дает возможность достаточно просто настраивать эти процедуры для обработки случаев терминоупотреблений, характерных как для решаемой прикладной задачи, так и для текстов конкретной ПО. Настройка осуществляется путем корректировки существующих и добавления новых шаблонов для терминологических словосочетаний, их вариантов и конструкций их употребления.

Апробация. Результаты диссертации докладывались:

– на международном семинаре Диалог по компьютерной лингвистике и ее приложениям в 2000 г. (Протвино, 2000) и в 2001 г. (Аксаково, 2001);

– на международной конференции Диалог по компьютерной лингвистике и интеллектуальным технологиям в 2004 г. (Верхневолжский, 2004), в 2007 г. (Бекасово, 2007) и в 2010 г. (Бекасово, 2010);

– на девятой, десятой и одиннадцатой национальных конференциях по искусственному интеллекту с международным участием КИИ-2004 (Тверь, 2004), КИИ-2006 (Обнинск, 2006) и КИИ-2008 (Дубна, 2008);

– на международной научной конференции студентов, аспирантов и молодых ученых Ломоносов, секция «Вычислительная математика и кибернетика» в 2008 г. (Москва, 2008) и в 2010 г. (Москва, 2010);

– на научно-исследовательском семинаре по методам построения программных систем (Москва, факультет ВМК МГУ, 2008);

– на научно-исследовательском семинаре «Динамические интеллектуальные системы» (Институт системного анализа РАН, 2009).

– на Ломоносовских чтениях: научной конференции, посвященной 300-летию со дня рождения М.В. Ломоносова (Москва, факультет ВМК МГУ, 2011).

Публикации. По теме диссертации опубликовано 13 работ, в том числе одна в издании, рекомендованном ВАК.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы и пяти приложений. Объем диссертации без приложений – 109 страницы, объем приложений – 16 страниц. Список литературы содержит 85 наименования.

Работа выполнена при частичной финансовой поддержке Минобрнауки России по государственному контракту от 16.05.2012 г. № 07.524.12.4018 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

Результаты проведенных исследований использовались в работе по гранту РФФИ № 06-01-00571 «Методы и средства интеллектуальной автоматической обработки текстов русскоязычных научно-технических документов».

СОДЕРЖАНИЕ РАБОТЫ

Во введении раскрывается тема диссертации, показываются ее научная новизна и актуальность, кратко описывается содержание работы по главам.

В главе 1 приведен обзор существующих методов извлечения терминологических слов и словосочетаний из текстов на естественном языке.

Методы извлечения терминов опираются на статистические и лингвистические критерии. Статистические критерии используют различные статистические величины, основанные на частоте встречаемости слов и словосочетаний и использующие предположение о том, что термины имеют тенденцию к многократному употреблению в тексте (или коллекции текстов). Выделенные по этим критериям слова и словосочетания рассматриваются как потенциальные термины.

Лингвистические критерии в первую очередь учитывают то, что термины, как правило, представляют собой именные словосочетания с определенной структурой, которую обычно описывают в виде **синтаксического образца** – он задает части речи составляющих термин слов и синтаксические связи между ними. К примеру, `N` – это образец, описывающий однословные термины-существительные (*ландшафт*, *аорта*), `A N` – образец терминологических словосочетаний из согласованных между собой прилагательного и существительного (*красное смещение*, *существенный пример*) и др. Если некоторое словосочетание текста имеет рассматриваемую структуру, то в системах автоматического извлечения терминов оно предположительно считается термином.

Другая лингвистическая информация, используемая для извлечения терминологических слов и словосочетаний, учитывает их употребление в рамках некоторых языковых конструкций (контекстов). К примеру, в текстах

сельскохозяйственной тематики из выявленных конструкции вида *yields of* SPECIES (*yield of wheat*, *yield of rice – урожай пшеницы*, *урожай риса*) извлекаются SPECIES – названия выращиваемых культур.

Один из основных недостатков методов извлечения терминов на базе статистических и лингвистических критериев связан с тем, что этим критериям удовлетворяют слова и словосочетания общеупотребительной лексики, (например: *задача*, *основная идея*, *применение правила*), т.к. они могут быть достаточно частотными и иметь типичную для терминов синтаксическую структуру. Как следствие, современные методы автоматического извлечения терминов позволяют получать в результате своей работы всего лишь **термины-кандидаты**, т.е. такие слова и словосочетания, для которых с той или иной степенью точности можно утверждать, что они являются терминами.

Повышение точности извлечения терминов можно достичь путем привлечения дополнительной статистической и лингвистической информации. Например, из множества терминов-кандидатов удаляются такие, которые входят в заранее составленный список стоп-слов. Оставшиеся кандидаты упорядочиваются на основе значений некоторой функции, и из упорядоченного списка отбираются слова и словосочетания, значения функции для которых выше или ниже установленного порога.

Основным способом повышения полноты извлечения терминов является учет текстовых вариантов, возникающих при употреблении терминов в тексте: варианты принимаются во внимание при подсчете частоты вхождения терминов. В настоящий момент существует два основных подхода к их автоматическому выявлению.

В рамках первого подхода осуществляется сравнение двух произвольных слов/словосочетаний, рассматриваемых либо как последовательности символов, либо как последовательности слов. Данный подход хорошо применим для распознавания близких по написанию терминологических вариантов вида *туннель – тоннель*, *colour of hammers – colour of this hammer*. В рамках второго подхода в формальном виде описываются правила образования текстовых вариантов, что обеспечивает выявление бóльшего количества различных видов вариантов (включая и соединения нескольких многословных терминов). Однако данный подход является языковозависимым, поскольку в целом правила варьирования терминов зависят от конкретного ЕЯ.

Проведенный обзор показывает, что для повышения полноты и точности извлечения терминов из НТ-текста целесообразно использовать определенную

комбинацию лингвистических и статистических критериев. При решении задачи извлечения терминологической информации из отдельного НТ-текста, рассматриваемой в данной диссертации, упор сделан на применение лингвистических критериев. Для формализации разнообразной лингвистической информации об употреблении терминов в НТ-текстах предложено использовать лексико-синтаксические шаблоны.

В главе 2 обсуждаются особенности терминов, конструкций и вариантов их употребления в НТ-текстах, вводится и характеризуется понятие лексико-синтаксического шаблона. Также описывается формализация типичных для научно-технической прозы языковых конструкций с терминами.

В зависимости от того, представлен или нет конкретный термин в компьютерном терминологическом словаре, используемом для обработки текста, будем называть этот термин соответственно **словарным** или **несловарным**.

Как правило, в НТ-текстах несловарные термины явно определяются или поясняются. К примеру, фраза

Периодическим расписанием цикла называется отображение T… (1)

определяет термин *периодическое расписание цикла*.

Несловарные термины, в противоположность словарным, обычно называют только формирующиеся понятия, их языковая форма выражения еще не устоялась, и поэтому для несловарных терминов характерно использование большого числа синонимичных названий. Такие синонимичные названия достаточно часто вводятся в тексте в рамках определенных конструкций. Например, во фразе

Назовем эти образования гипержанрами, или гипержанровыми формами

вводится несловарный термин *гипержанр* и его синоним *гипержанровая форма*.

Важной особенностью как словарных, так и несловарных терминов является употребление в НТ-тексте их текстовых вариантов. В рамках одного текста характерно употребление **лексико-синтаксических вариантов** терминов, когда изменяется лексический состав и синтаксические связи составляющих термин слов (*аэробное упражнение – упражнение, дисковый контроллер – контроллер диска*), и **вариантов сокращения** (*саморегулируемые организации – СРО, мономолекулярный слой – монослой*).

Кроме вариантов отдельного термина в НТ-текстах наблюдаются **соединения нескольких терминов**, при образовании которых термины часто разрываются, а их общие части сокращаются, что затрудняет автоматическое извлечение их из текста. Например, соединение *разрядность внутренних регистров* построено на основе двух терминов: *разрядность регистра* и *внутренний регистр*, в нем наблюдается разрыв термина *разрядность регистра* и слияние общей части – слова *регистр*.

Формализация рассмотренных конструкций с терминами проводилась на базе лексико-синтаксических шаблонов. **Лексико-синтаксический шаблон** обычно состоит из **имени** и **тела**. Имя записывается перед телом через знак равенства; если шаблон не будет использоваться в других шаблонах, то имя можно не указывать.

Тело определяет последовательность **элементов**, из которых должна состоять описываемая языковая конструкция, и задает условия синтаксического согласования этих элементов. К примеру, фразы вида (1) описываются следующим лексико-синтаксическим шаблоном:

```
T1<c=ins> V<называться,t=pres,p=3,m=ind> D1<c=nom>
                    <T1.n=V.n> # T1   ,
```

где `V<называться,t=pres,p=3,m=ind>` – элемент-слово, описывающее две словоформы: *называется* и *называются*;
`T1` и `D1` – экземпляры ранее определенных шаблонов с именами `T` и `D` (эти шаблоны задают соответственно синтаксическую структуру терминов и определяющих эти термины фраз);
`<T1.n=V.n>` – условия согласования грамматического числа определения `T1` и глагола `V`.

Правила извлечения самих терминов из распознанных по шаблону конструкций задаются с помощью **шаблонов извлечения**, которые записываются после тела шаблона (за символом #). Так, из фразы (1) с помощью шаблона извлечения `T1` будет выделен термин *периодическое расписание цикла*.

Средства лексико-синтаксических шаблонов были применены для формализации структуры терминологических слов и словосочетаний, а также конструкций и вариантов их употребления. Примеры полученных лексико-синтаксических шаблонов приведены в Таблице 1 (шаблон `AP = A|Pa` задает понятие адъектива, т.е. прилагательного или причастия).

Таблица 1. Примеры лексико-синтаксических шаблонов

Тип употребления	Шаблон	Примеры конструкций
Словарные термины	N1<критерий>["Рэлея"| N2<подобие,c=gen>]	*критерий*, *критерий Рэлея*, *критерий подобия*
Синтаксические образцы терминов	SP = N1 | N1 N2<c=gen> | AP1 AP2 N1 <AP1=AP2=N1>| AP1 N1 <AP1=N1>	*вектор*, *генератор шума*, *жесткий магнитный диск*, *альбедный мюон*
Конструкции определений терминов	Term1<c=nom> ["–"] "это" Defin1<c=nom> # Term1	*Информационный ресурс – это набор текстовых файлов…*
Конструкции введения синонимов	SP1 "(" SP2 ")"<SP1.c= SP2.c> # SP1, SP2	*…создание информационных систем (ИС)…*
Соединения терминов	A1 A2 N1 <A1=A2=N1> # A1 N1 <A1=N1>, A2 N2 <A2=N2>	*длинных целых чисел – длинное число, целое число*
Правила образования вариантов	N1 N2<c=gen> # N1, N3 N2<c=gen> <Syn(N1,N3)>	*метка адреса – метка, маркер адреса*

Созданные наборы лексико-синтаксических шаблонов составляют базу лингвистической информации, на которую опираются разработанные в данной диссертации процедуры извлечения из НТ-текста на русском языке различных употреблений терминов. Работа процедур сводится к наложению лексико-синтаксических шаблонов на обрабатываемый текст, поиску в нем текстовых фрагментов, описываемых шаблонами, и извлечению из этих фрагментов соответствующих терминоупотреблений.

В главе 3 подробно описываются разработанные процедуры извлечения из НТ-текста терминологических слов и словосочетаний, а также конструкций и вариантов употребления терминов на основе их формального описания в виде лексико-синтаксических шаблонов. В конце главы охарактеризована программная реализация процедур извлечения.

В виде лексико-синтаксических шаблонов формализована разнотипная лингвистическая информация: о распознаваемых терминах, их вариантах и конструкциях их употребления. Для каждого типа информации разработаны соответствующие процедуры извлечения. Процедуры образуют 3 группы:

1) Процедуры извлечения вхождений терминов:

- **`getDictTerms`** – извлечение словарных терминов;

- **getNonDictTerms** – извлечение несловарных терминов.

2) Процедуры распознавания языковых конструкций с последующим извлечением из них терминов:

- **getAuthTerms** – извлечение авторских терминов;
- **getSynTerms** – извлечение синонимов терминов;
- **getTermsfromCombs** – извлечение терминов из соединений.

3) Процедура распознавания текстовых вариантов терминов: **getVarsforTerms**.

Процедура **getDictTerms** работает с шаблонами словарных терминов, а процедура **getNonDictTerms** – с шаблоном, описывающим типичные синтаксические образцы терминов. В результате наложения шаблонов на текст выделяются текстовые фрагменты, соответствующие вхождениям распознанных терминов в текст.

Процедуры **getAuthTerms**, **getSynTerms** и **getTermsfromCombs** получают на вход шаблоны, тела которых описывают распознаваемые языковые конструкции, а шаблоны извлечения задают правила извлечения терминов из этих конструкций. В данном случае наложение на текст тела шаблона дает текстовые фрагменты, представляющие вхождения в текст распознанных конструкций, из которых затем извлекаются сами термины.

Процедура **getVarsforTerms** работает с шаблонами, тела которых описывают синтаксические образцы терминов, а шаблоны извлечения задают синтаксические образцы возможных текстовых вариантов. В этом случае наложение тела шаблона позволяет получать текстовые фрагменты, соответствующие распознанным терминам, на основе которых затем формируются их варианты. Сформированные варианты используются процедурой для выявления их вхождений в обрабатываемый текст.

Указанные процедуры извлечения терминов и их различных употреблений реализуют поверхностный синтаксический анализ и применяются к тексту T, который после графематического и морфологического анализа представляет собой последовательность простых фрагментов текста t_i – словоформ и разделяющих их символов: $T=(t_1, t_2, \ldots, t_{nT})$, расположенных в том же порядке, что и в исходном тексте. Для каждого простого фрагмента t_i, являющегося словоформой, известна часть речи этого слова, начальная форма и набор значений его морфологических характеристик χ_i (рода, числа, времени и т.д.). Общая схема обработки текста представлена на Рисунке 1.

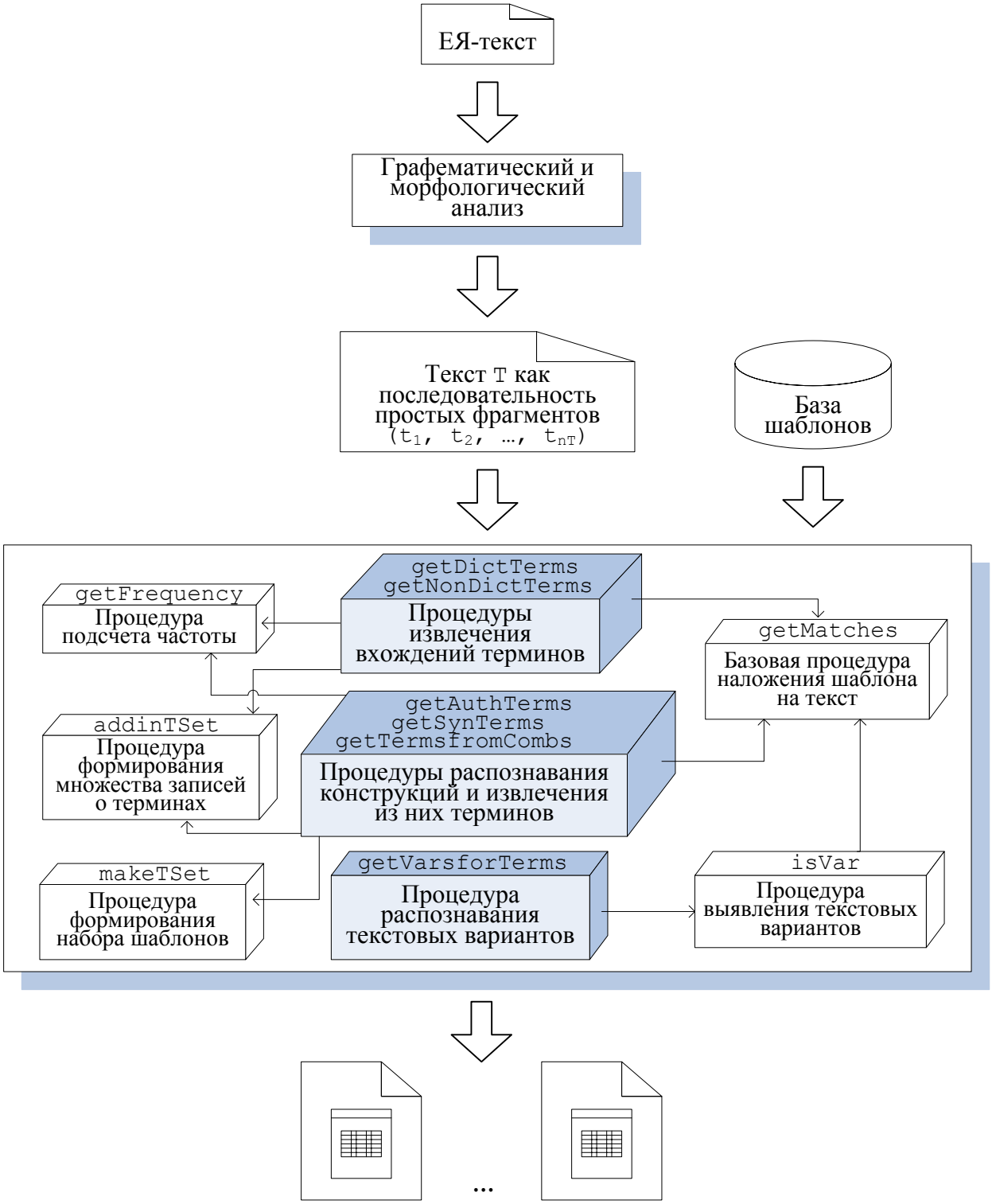

Рисунок 3.1. Общая схема обработки текста

На вход каждой из описываемых процедур извлечения поступает анализируемый текст $T=(t_1, t_2, \ldots, t_{nT})$ и набор шаблонов $S=\{P_1, P_2, \ldots, P_{nS}\}$. Работа любой из этих процедур заключается в наложении каждого шаблона P_i из набора S на текст T и последующей обработке полученных при этом результатов; на выходе процедуры – множество $MR=\{mr_1, mr_2, \ldots, mr_{nM}\}$ извлеченных терминах-кандидатах с информацией о частоте употребления каждого из них.

В своей работе процедуры извлечения опираются на базовую процедуру `getMatches`, отвечающую за наложение шаблона на текст, и следующие вспомогательные процедуры:

- `getFrequency` – подсчет частоты употребления терминов в тексте с учетом вложений терминологических словосочетаний друг в друга;
- `addinTSet` – формирование множества записей об извлеченных терминах;
- `makePSet` – построение дополнительного набора шаблонов (для последующего наложения их на текст);
- `isVar` – выявление текстовых вариантов терминов.

Реализация всех разработанных процедур проводилась на языке C++ с использованием библиотеки LSPL[2]; разработка проводилась под ОС Linux. Общий объем базы лексико-синтаксических шаблонов составил около 6500 шаблонов.

Каждая из процедур извлечения позволяет распознавать в тексте и обрабатывать определенный тип терминоупотребления. При этом получаемые процедурами множества терминов-кандидатов в общем случае пересекаются, поскольку одно и тоже слово или словосочетание может быть выявлено разными процедурами. Таким образом, для решения задачи автоматического извлечения терминов из конкретного НТ-текста требуется объединение результатов работы этих процедур.

В главе 4 описывается экспериментальное исследование эффективности работы реализованных процедур и формулируется стратегия объединения их результатов, цель которой – увеличение F-меры, вычисляемой как гармоническое среднее полноты и точности извлечения. В конце главы

[2] Большакова Е.И., Носков А.А. Программные средства анализа текста на основе лексико-синтаксических шаблонов языка LSPL // Программные системы и инструменты: Тематический сборник, №11 / Под ред. Королева Л.Н. – М.: Изд. отдел факультета ВМК МГУ имени М.В. Ломоносова; МАКС Пресс, 2010, с. 61-73.

продемонстрированы пути применения разработанных процедур в прикладных задачах АОТ, в которых требуется проведение терминологического анализа отдельно взятого НТ-текста. В качестве таких задач взяты составление глоссария и предметного указателя научно-технического документа.

Исследование разработанных процедур извлечения употреблений терминов проводилось на коллекции научно-технических текстов из двух предметных областей – информатика и вычислительная техника (ИиВТ) и физика. Оценка результатов работы процедур происходила путем сравнения множеств терминов-кандидатов, полученных каждой процедурой, с эталонными множествами терминов, сформированными экспертами.

По результатам сравнения были выявлены причины снижения полноты и точности извлечения терминов и их употреблений. Так, основная причина снижения полноты во всех процедурах связана с особенностью языковых конструкций ЕЯ. В частности, отдельные термины и соединения нескольких терминологических словосочетаний могут иметь сходную структуру: например, термин *число большой разрядности* и соединение *выделение динамической памяти* (состоящее из словарных терминов *выделение памяти* и *динамическая память*). При обработке обоих словосочетаний как соединений будет потерян термин *число большой разрядности*, а при их обработке как отдельных терминов не будут выделены термины из соединения.

Основной же причиной снижения точности являются ограничения используемых лингвистических критериев. Например, типичным синтаксическим образцам терминов в тексте соответствует большое количество слов и словосочетаний, не являющихся терминами – *способ, малая часть, решение поставленной задачи*. Тем самым, множества терминов-кандидатов, полученные процедурами извлечения, следует обработать дополнительно для исключения из них подобных слов и словосочетаний. В частности, при обработке результатов извлечения несловарных терминов для повышения точности извлечения в дополнение к лингвистическим критериям предложено использовать статистическую характеристику – среднее взвешенное арифметическое всех частот несловарных терминов-кандидатов:

$$F = \frac{\sum_i f_i}{\sum_i n_i} \quad ,$$

где f_i – значение частоты употребления i-ого термина-кандидата, а знаменатель дроби представляет собой количество разных слов и словосочетаний, выявленных процедурой. Кандидаты, частота употребления

которых ниже округленного значения F, при прочих равных не считаются терминами.

Согласно предложенной в диссертации стратегии сначала к рассматриваемому тексту по отдельности применяются разработанные процедуры извлечения терминоупотреблений, и в результате их применения получаются множества терминов-кандидатов. Затем из этих множеств по эвристическим правилам, сформулированным по итогам проведенного экспериментального исследования, отбираются наиболее вероятные кандидаты. Правила применяются по очереди, и в результате итерационно строится итоговое множество M записей об отобранных терминах-кандидатах и итоговый набор групп текстовых вариантов G_1, G_2, ..., G_N.

Правила делятся на три группы:

1) Правила начального формирования множества M (3 правила).

2) Правила расширения множества M за счет учета вариантов употребления терминов (8 правил).

3) Правило формирования групп текстовых вариантов G_1, G_2, ..., G_N (1 правило).

Предложенная стратегия была применена к коллекции научно-технических текстов для извлечения из них терминов и их различных употреблений. Полученные результаты были сопоставлены с результатами работы наиболее известных и часто используемых методов извлечения терминов, полученных для этой же тестовой коллекции. По сравнению с одним из этих методов – методом Terms--[1], дающим наилучшие результаты извлечения терминов и их употреблений, F-мера извлечения терминов увеличилась на 17,6%, а F-мера извлечения всех их употреблений – на 11,7%; для правильно извлеченных терминов полнота распознавания их различных употреблений выросла на 5,3%.

В **заключении** сформулированы основные результаты диссертационной работы, выносимые на защиту:

1. Предложен подход, позволяющий формализовать в виде лексико-синтаксических шаблонов структуру терминологических словосочетаний, а также конструкции и варианты их употребления для использования в процедурах автоматического извлечения из текста терминологической информации.

2. В рамках предлагаемого подхода разработаны процедуры извлечения из текста терминологической информации, опирающиеся на ее формальное

описание в виде шаблонов и допускающие настройку за счет изменения используемого набора шаблонов.

3. Разработанные процедуры программно реализованы, проведено их экспериментальное исследование на базе созданного набора шаблонов терминов, их вариантов и конструкций их употребления.

4. По результатам экспериментального исследования предложена стратегия объединения результатов работы реализованных процедур, позволяющая в целом улучшить показатели точности и полноты извлечения терминов из текста.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Большакова Е.И., Васильева Н.Э. К вопросу об автоматизации литературно-научного редактирования // Компьютерная лингвистика и ее приложения: Труды Международного семинара Диалог‘2000. – Протвино, 2000. – Т.2. – С. 59-63.

2. Большакова Е.И., Васильева Н.Э., Юдин Д.А. Выделение словарных терминологических словосочетаний в научно-технических текстах // Компьютерная лингвистика и ее приложения: Труды Международного семинара Диалог‘2001. – Аксаково, 2001. – с. 48-51.

3. Васильева Н.Э. Шаблоны употреблений терминов и их использование при автоматической обработке научно-технических текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог‘2004. – М.: Изд-во РГГУ, 2004. – Т. 2. – С. 96-101.

4. Большакова Е.И., Баева Н.В., Васильева Н.Э. Структурирование и извлечение знаний, представленных в научных текстах // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. – М.: Физматлит, 2004. – Т. 2. – С. 480-488.

5. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. – М.: Физматлит, 2006. – Т. 2. – С. 506-524.

6. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные

технологии: Труды Международной конференции Диалог‘2007. – М.: Изд-во РГГУ, 2007. – Т. 2. – С. 70-75.

7. Васильева Н.Э. Распознавание в научно-технических текстах терминов и их вариантов // Ломоносов – 2008: Материалы XV Международной научной конференции студентов, аспирантов и молодых ученых: секция «Вычислительная математика и кибернетика». Сборник тезисов. – 2008. – С. 23.

8. Большакова Е.И., Васильева Н.Э. Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ-2008. Труды конференции в 3-х томах. – М.: Физматлит, 2008. – Т. 2. – С.174-182.

9. Большакова Е.И., Васильева Н.Э. Формализация лексико-синтаксической информации для распознавания регулярных конструкций естественного языка // Программные продукты и системы. – 2008. – № 4. – С. 103-106.

10. Антонов В.Ю., Ефремова Н.Э. Автоматическое выявление терминологических вариантов в русскоязычных текстах // Ломоносов – 2010: Материалы XVII Международной научной конференции студентов, аспирантов и молодых ученых: секция «Вычислительная математика и кибернетика». Сборник тезисов. – 2010. – С. 80.

11. Ефремова Н.Э., Большакова Е.И., Носков А.А., Антонов В.Ю. Терминологический анализ текста на основе лексико-синтаксических шаблонов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог‘2010. – М.: Изд-во РГГУ, 2010. – С. 124-129.

12. Bolshakova E., Efremova N., Noskov A. LSPL-Patterns as a Tool for Information Extraction from Natural Language Texts // K.Markov et al. (eds.): New Trends in Classification and Data Mining, ITHEA. – 2010. – P. 110-118.

13. Большакова Е.И., Ефремова Н.Э., Носков А.А. Методы и средства построения программных систем для анализа текста с использованием лингвистических шаблонов // Ломоносовские чтения: научная конференция, посвященная 300-летию со дня рождения М.В. Ломоносова: Тезисы докладов. – 2011. – С. 97.