На правах рукописи

СИМАКОВ Константин Васильевич

МОДЕЛИ И МЕТОДЫ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Научный руководитель –
доктор технических наук
профессор В.Н. Голубкин

Москва – 2008

Работа выполнена на кафедре «Компьютерные системы и сети» (ИУ-6) факультета «Информатика и системы управления» (ИУ) Государственного образовательного учреждения высшего профессионального образования «Московский государственный технический университет имени Н.Э. Баумана» (МГТУ имени Н.Э. Баумана).

Научный руководитель:	доктор технических наук, профессор Голубкин Виктор Николаевич
Официальные оппоненты:	доктор физико-математических наук, профессор Осипов Геннадий Семенович
	кандидат технических наук Шабанов Владислав Игоревич
Ведущая организация:	ЗАО «Концерн ВНИИНС» (Всероссийский научно-исследовательский институт автоматизации управления в непромышленной сфере)

Защита состоится 13 марта 2008 г. в 16 часов 00 минут на заседании диссертационного совета Д.212.141.10 по защите диссертаций при Московском государственном техническом университете имени Н.Э. Баумана по адресу: 105005, г. Москва, ул. 2-я Бауманская, д. 5.

С диссертацией можно ознакомиться в библиотеке Московского государственного технического университета имени Н.Э. Баумана по адресу: 105005, г. Москва, ул. 2-я Бауманская, д. 5.

Автореферат разослан 1 февраля 2008 г.

Ученый секретарь
Диссертационного совета Д.212.141.10
кандидат технических наук, доцент Иванов С.Р.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Наиболее распространенной формой представления знаний являются естественно-языковые тексты. Текстовая форма знаний естественна для человека, такие знания легко воспринимаются, порождаются, тиражируются и модифицируются. Однако интенсивный рост текстовых массивов является причиной трудной доступности целевых знаний, когда в них возникает потребность. Дополнительной проблемой является сложность валидации текстовых массивов, состоящей в поиске и исправлении ошибок, устранении дубликатов и противоречий. Информационно-поисковые системы не рассчитаны на решение данной задачи, поскольку оперируют словами текста, а не знаниями, содержащимися в нем.

В связи с этим приобретают актуальность системы извлечения знаний из текстов. В результате извлечения знания приобретают явный вид и становятся пригодными для автоматизированной обработки, например, системами сопоставляющего анализа, выполняющими сопоставление результата извлечения с эталонной моделью предметной области с целью его валидации.

Проблеме извлечения посвящено множество зарубежных работ, объединяемых в единый класс задач извлечения информации из текстов. Извлекаемая информация представлена структурами данных, поля которых заполняются текстовыми фрагментами. Недостатком зарубежных разработок является сильная зависимость от конкретной грамматики языка. Среди отечественных работ известны только две законченные системы компаний RCO и Yandex, имеющие крайне ограниченное применение, поскольку не существует простого способа их адаптации к произвольной предметной области. Более того, в современных работах нет сведений о системах сопоставляющего анализа, находящихся в эксплуатации.

Таким образом, разработка математической модели извлечения, применимой для обработки текстов без привязки к конкретному языку и легко адаптируемой под нужды конкретной предметной области, представляет собой важную научную задачу, а разработка модели представления знаний, в рамках которой формируется результат извлечения, удобный для выполнения сопоставляющего анализа, имеет существенное практическое значение.

Извлечение информации из текстов является подзадачей более крупной задачи, решению которой посвящена диссертация, а именно - извлечению знаний. Чтобы выявлять в текстах структуры данных, необходимо располагать двумя наборами правил: правилами морфологического анализа и правилами извлечения. Первые выявляют лингвистические свойства слов текстов, тогда как вторые, используя эти свойства, накладывают условия на состав и структуру контекстов целевой информации.

Правила обоих типов наравне с извлекаемыми структурами данных являются знаниями предметной области. Формирование таких правил в существующих отечественных разработках осуществляется вручную, что является причиной сложности настройки системы извлечения.

В связи с этим разработка методов автоматизированного составления правил извлечения и правил морфологического анализа является актуаль-

ной задачей, решение которой в общем виде без привязки к конкретному языку в настоящий момент отсутствует.

Цель и основные задачи работы. Целью работы является разработка моделей извлечения знаний из текстов и методов их обучения для систем сопоставляющего анализа текстов на естественном языке. Для достижения поставленной цели в рамках диссертации решены следующие задачи:

1. исследование современных моделей извлечения информации из текстов и методов обучения таких моделей;
2. разработка модели представления знаний, позволяющей эффективно выполнять сопоставляющий анализ текстов;
3. создание модели извлечения знаний из предметно-ориентированных текстов;
4. разработка метода обучения модели извлечения знаний из текстов;
5. создание модели морфологического анализа слов и метода ее обучения;
6. экспериментальная проверка предложенных моделей и методов.

Объект и предмет исследования. Объектом исследования являются естественно-языковые тексты как форма представления знаний предметной области. Предметом исследования являются процессы автоматизированного выявления и формализации знаний, представленных в форме естественно-языковых текстов.

Методы исследования. Теоретические исследования проведены с использованием теоретико-множественного аппарата, методов теории вероятностей и теории информации, аппарата логических исчислений. При разработке моделей и методов был применен аппарат алгебраических систем, в том числе алгебраических решеток и графов, а также аппарат формальных грамматик и теория автоматов.

Научная новизна работы заключается в следующем.

1. Предложена модель извлечения фреймовых слотов из предметно-ориентированных текстов. Введенная в модели решетка лексических ограничений позволила теоретически обосновать возможность обучения модели. Простота структуры правил извлечения обеспечивает практическую реализуемость механизмов машинного обучения, а также реализацию метода извлечения на основе конечного автомата, независимого от грамматики естественного языка.
2. Разработан метод обучения модели извлечения, в рамках которого предложена новая сжимающая стратегия группового обобщения обучающих примеров, а также новый подход к парному обобщению правил на основе оценки совокупной погрешности обобщения их отдельных элементов.
3. Предложена модификация принципа аналогии морфологического анализа текстов, позволяющая существенно сократить объем морфологического словаря и снизить вычислительную сложность алгоритма анализа.
4. Разработана модель морфологического анализа, действующего в соответствии с модифицированным принципом, а также предложен метод ее обучения, позволяющий без вмешательства человека построить морфологический анализатор, обладающий лучшим качеством анализа в сравнении со словарными методами.

Достоверность научных положений и выводов диссертационной работы подтверждена практической реализацией разработанных моделей и методов, результатами проведенных экспериментов, а также внедрением и опытной эксплуатацией в ряде систем сопоставляющего анализа текстов.

Практическая ценность. Разработанная модель извлечения может быть использована в рамках систем сопоставляющего анализа, выполняющих поиск орфографических ошибок, восстановление пропущенных в тексте данных, а также выявление противоречий между содержимым текста и эталонной базой знаний. Модель применима в системах извлечения информации из текстов в следующих областях: автоматизированное наполнение реляционных баз данных, справочников, словарей и онтологий, информационная разведка, мониторинг текстовых потоков.

Реализация модели извлечения использовалась при разработке:

- «Системы семантического контроля текстов редактируемых документов», применяемой для выявления несоответствий в текстах стенограмм заседаний Совета Федерации Федерального Собрания Российской Федерации.
- «Интеллектуальной системы выявления и исправления ошибок в почтовых адресах клиентов банка», применяемой для валидации российских почтовых адресов, представленных в текстовой форме.

Разработанный метод обучения модели извлечения избавляет эксперта от ручного формирования правил извлечения.

Модифицированный принцип аналогии морфологического анализа и построенная на его основе модель позволяет на порядок сократить объем морфологического словаря. Предложенный метод обучения данной модели полностью избавляет эксперта от подготовки обучающих примеров.

Кроме приведенных выше систем модель морфологического анализа и метод ее обучения встроены в следующие программные комплексы:

- Информационно-поисковая система «Обзор СМИ», эксплуатируемая в Совете Федерации РФ, в рамках которой встроен модуль морфологического анализа для задачи автоматического построения аннотаций.
- Комплекс программных средств специального назначения по классификации и кластеризации текстов документов, разработанный и использованный в рамках НИР, выполненной ВНИИНС для МО.

Апробация результатов работы. Результаты проведенных в диссертационной работе исследований опубликованы в 11 работах, из них в журналах из перечня ВАК - 2 статьи. Основные положения работы докладывались и обсуждались:

- на 5-ой, 6-ой, 7-ой, 8-ой и 9-ой всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» в 2003, 2004, 2005, 2006 и 2007 годах;
- на научной конференции «Информатика и системы управления в XXI веке» в 2003 году на факультете «Информатика и системы управления» МГТУ им. Н.Э. Баумана.

Личный вклад автора. Все основные научные результаты, модель извлечения, метод ее обучения, модель морфологического анализа, действующего в соответствии с предложенной модификацией принципа аналогии,

метод обучения данной модели, разработанные на их основе алгоритмы и программные средства, экспериментальные исследования, приведенные в диссертации, получены автором лично.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из семи глав, введения, заключения, списка литературы и двух приложений. Общий объем диссертации – 267 страниц, включая 88 рисунков и 24 таблицы. Библиография включает 116 наименований, из которых 83 иностранных источника.

Во введении обосновывается актуальность темы диссертации. Формулируется цель и основные задачи работы, характеризуется ее научная новизна.

В главе 1 «Постановка задачи извлечения знаний из текстов» дается определение решаемым в диссертации задачам.

В процессе сопоставления текст подлежит стадийной обработке. Каждый функциональный блок использует некоторые знания о предметной области, начиная с сегментатора, разбивающего текст на отдельные составляющие, и заканчивая компаратором, реализующим логику сопоставления извлеченных знаний. Для качественного функционирования каждого блока необходимо обеспечить полноту и актуальность используемых знаний. Для этих целей служит система извлечения.

В главе приводится классификация знаний, пополнение которых целесообразно выполнять с минимальным участием эксперта. К ним отнесены процедурные знания о языке предметной области (правила словообразования и согласования слов), экземпляры декларативных знаний предметной области и правила их выявления. На рис. 1 приведена функциональная схема системы извлечения этих знаний.

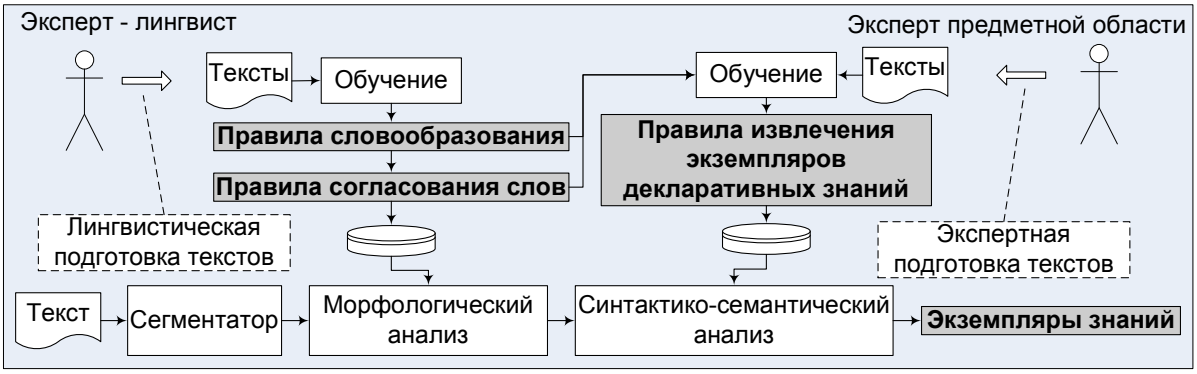

Рис. 1. Функционирование системы извлечения

Темными блоками выделены знания, извлекаемые из текстов на каждом из трех крупных этапов: формирование правил о языке, формирование правил извлечения и непосредственное извлечение экземпляров. На первом этапе человек-эксперт по лингвистике подбирает типовые тексты предметной области. Эти тексты пропускаются через процедуру автоматического обучения, которая формирует правила словообразования и согласования слов. На втором этапе другой человек-эксперт предметной области выполняет подготовку другого массива текстов. Эта подготовка подразумевает явное выделение в текстах экземпляров декларативных знаний предметной области.

Такое явное выделение выполняется в виде аннотирования или разметки. Размеченные тексты поступают на вход второй процедуре обучения, которая, используя полученные знания о языке, формирует правила извлечения. На третьем этапе другая выборка текстов, в которой, отражены целевые экземпляры, подвергается морфологической и синтактико-семантической обработке с применением правил, созданных на первых двух этапах. Результатом этого являются извлеченные экземпляры декларативной части знаний.

Таким образом, необходимо разработать две математические модели, закладываемые в основу функциональных блоков «морфологический анализ» и «синтактико-семантический анализ», а также методы их обучения.

В главе 2 «Обзор методов извлечения знаний» проведено исследование современного положения дел в области машинного обучения и извлечения информации из текстов.

В области машинного обучения выделено два класса методов, основывающихся на объяснении и на подобии, второй класс отмечен, как наиболее подходящий для поставленных задач.

В области извлечения информации из текстов рассмотрены два подхода: символьный и вероятностный. В рамках первого проанализированы 11 обучаемых зарубежных систем извлечения, оперирующих правилами в виде регулярных выражений. Из недостатков этих систем выделены следующие.

- Нет ограничений на итерацию элементов правил снизу и сверху.
- Ни одна из существующих систем не позволяет налагать запреты на употребление в тексте отдельных слов или их классов, заданных признаками.
- Применяемые методы обучения реализуют покрывающую стратегию, возможности сжимающей стратегии исследованы слабо.
- В основном используется формирование правила на основе одного единственного обучающего примера, групповое обобщение не реализовано ни в одной из систем.

В рамках вероятностного подхода выделено шесть направлений, наиболее перспективными из которых являются модели, максимизирующие энтропию, и условные случайные поля. Вероятностные методы не позволяют предсказывать влияние изменений модели на качество ее извлечения после обучения. Кроме того, вероятностные методы не позволяют выполнять анализ полученных в результате обучения знаний, поскольку эти знания представлены в виде распределений вероятностей, неудобных для восприятия экспертом. С учетом этого выделены следующие предпосылки разработок.

1. В настоящий момент не существует обучаемых моделей извлечения знаний из текстов, не привязанных к конкретному языку, обладающих должным качеством извлечения.
2. Необходимо иметь возможность как точной ручной настройки модели, так и автоматической – на основе обучающих примеров. Данному требованию в большей степени удовлетворяют символьные модели.
3. Модель извлечения должна оперировать с неограниченным числом атомарных признаков, приписываемых словам текста (в том числе морфологических), и не должна привязываться к конкретному синтаксису.

4. Чтобы правила извлечения можно было использовать в качестве компонентов более сложных моделей, их синтаксис должен быть предельно простым. Должно быть минимизировано количество обязательных предварительных анализаторов.
5. Рассмотренные методы символьного обучения не используют сжимающую стратегию, которая потенциально способна выполнять более тщательный поиск целевых правил извлечения.

В главе 3 «Разработка принципов построения систем сопоставляющего анализа и извлечения знаний» предложена комбинированная модель представления знаний предметной области, сочетающая в себе достоинства онтологического и фреймового представлений. На этапе извлечения используется фреймовое представление, которое затем преобразуется в объекты онтологии с означенными свойствами и взаимосвязями. Это достигается за счет наложения схемы фреймов на схему онтологии, для обеспечения возможности наложения к схемам предъявлен ряд следующих требований.

1. $\forall c \in C\, \exists d \in D : (c,d) \in R_{CD}$, где C и D – множества классов и простых свойств классов онтологии, R_{CD} – отношение принадлежности свойств классам.
2. $\forall f \in F \exists s \in S\ : (f,s) \in R_{FS}$, где F – множество фреймов, S – множество слотов, R_{FS} – отношение принадлежности слотов фреймам.

Оба свойства необходимы, т.к. при извлечении экземпляра фрейма доступными являются лишь значения его слотов, без которых невозможно выявить наличие в тексте целевого экземпляра. Для обеспечения наложения схем сформулированы следующие необходимые и достаточные условия.

1. $\forall (f,s) \in R_{FS} : (s,T_i) \in R_{ST} \Rightarrow \exists !(c,d) \in R_{CD} \wedge \exists !(d,T_i) \in R_{DT}$, где R_{ST} – отношение, связывающее с каждым слотом s область его допустимых значений T_i, R_{CD} – аналогичное отношения для простых свойств классов.
2. $$\left.\begin{array}{l}\forall (f,s_1) \in R_{FS} : (s_1,T_i) \in R_{ST} \\ \forall (f,s_2) \in R_{FS} : (s_2,T_j) \in R_{ST}\end{array}\right\} \Rightarrow \begin{cases}\exists !(c_1,d_1) \in R_{CD} \wedge \exists !(d_1,T_i) \in R_{DT} \\ \exists !(c_2,d_2) \in R_{CD} \wedge \exists !(d_2,T_j) \in R_{DT} \\ c_1 \neq c_2 \Rightarrow \exists C_f(c_1,c_2)\end{cases},$$
   где $C_f(c_1,c_2)$ – цепь классов вида $c_1,...c_i,...c_2$ такая, что $\forall c_i,c_{i+1} \in C_f(c_1,c_2) \Rightarrow \begin{bmatrix}\exists l : (c_i,l) \in R_{CL} \wedge (l,c_{i+1}) \in R_{LC} \\ \exists l : (c_{i+1},l) \in R_{CL} \wedge (l,c_i) \in R_{LC}\end{bmatrix}$, где R_{CL} и R_{LC} – отношения инцидентности между связями и классами онтологии. Классы указанной цепи должны удовлетворять условию $\forall c_i \in C_f(c_1,c_2) \Rightarrow (f \to c_i)$, где $f \to c$ обозначает тот факт, что одним из классов, на которые наложен фрейм f, является класс c.

В главе предложена реализация сопоставляющего анализа с учетом комбинированного использования фреймового и онтологического представления. Приведена схема анализа, использованная в рамках внедрения результатов диссертации.

Для обучения морфологического анализатора и модели извлечения предложена единая стратегия обучения (см. рис. 2).

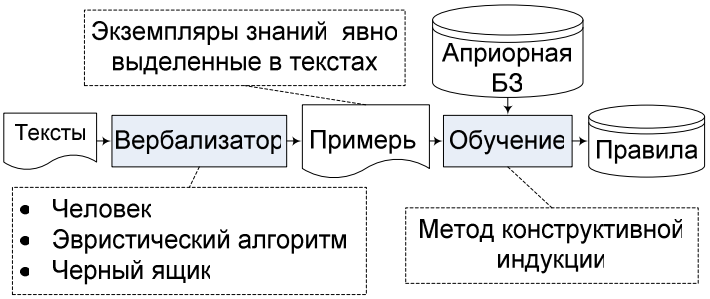

Рис. 2. Схематичное отражение стратегии обучения

Важным этапом обучения, в соответствии с данной стратегией, является вербализация (приведение к явному виду) знаний представленных в тексте. Для морфологического анализа вербализатором является сторонний анализатор, действующий на ограниченном лексиконе языка. Для модели извлечения экземпляров фреймов вербализацию может выполнять человек-эксперт. В обоих случаях результатом обучения являются правила, принципиально отличающиеся от тех, которыми руководствуется вербализатор при подготовке обучающих примеров.

Описанной стратегии присвоено название «Акцентированная Аппроксимирующая Абстракция» (сокращенно А3). Абстракция заключается в том, что в процессе обучения выполняется смена формулировки. Исходные знания представлены в виде неявных правил, которыми руководствовался автор при составлении текстов. Поскольку о формулировке этих правил ничего не известно, можно считать, что она не удовлетворительна и требуется ее замена. В главе 2 такая замена в рамках машинного обучения отнесена к классу дедуктивной абстракции. Аппроксимация обусловлена тем, что знания, полученные в результате обучения, не претендуют на точное отражение знаний, которыми руководствовался составитель текстов. Акцентированность заключается в том, что в результате обучения отражены только те знания, которые задействованы в выборке наблюдаемых текстов, хотя на самом деле вербализатор может располагать куда большими знаниями о текстах предметной области. Выборка наблюдений является своего рода акцентом, позволяющим выполнить вербализацию и обобщить только ту часть знаний, которая отражена в этой выборке.

В главе 4 «Разработка модели извлечения экземпляров фреймов» предложена сегментная модель текста TM, которой должен отвечать любой текст, подлежащий обработке, а также модель извлечения фреймовых слотов EM.

Модель текста определяется алгебраической системой вида $TM = < T, W, t_{\varnothing}, \bullet >$, позволяющей представлять произвольный текст в виде сцепленных сегментов, где T – множество текстовых сегментов, W – множество слов модели, $t_{\varnothing}$ - пустой текстовый сегмент, $\bullet$ – операция сцепления на T. Операция $\bullet$ позволяет из произвольной пары текстовых сегментов сформировать новый сегмент, $t_{\varnothing}$ является нейтральным элементом по отношению к операции сцепления.

Использование $t_{\varnothing}$ позволяет рассматривать текстовые сегменты разной длины в виде сцеплений из одного и того же числа n сегментов вида $t = t_1 \bullet t_2 \bullet ... \bullet t_n$, что определяет возможность построения текстовых шаблонов (образцов), являющихся основным компонентом правил извлечения. Мини-

мальными сегментами являются слова W, к которым также отнесены знаки препинания. Любой текстовый сегмент может быть представлен в виде сцепления слов.

Модель извлечения определяется как алгебраическая система вида $EM =< V, P, R, p_\varnothing, \circ, <_p >$, где V – множество правил извлечения, P – множество образцов текстовых сегментов, R – множество элементов образцов, $p_\varnothing$ – пустой образец, $\circ$ – операция сцепления образцов, $<_p$ – отношение порядка между элементами образца $p \in P$. Компоненты модели EM обладают следующими свойствами.

- $\forall p_1 \in P \wedge \forall p_2 \in P \; \exists p = p_1 \circ p_2$.
- $p_\varnothing \in P \wedge \forall p \in P \Rightarrow p = p_\varnothing \circ p \wedge p = p \circ p_\varnothing$.
- $\forall v \in V \Rightarrow v = p_b \circ p_c \circ p_a$ – правила извлечения представляют собой сцепление тройки образцов: префиксный p_b, извлекающий p_c и постфиксный p_a.

Следствия приведенных свойств модели:

- $\forall p \in P \Rightarrow p \in V$ – образец может быть представлен в виде правила.
- $\forall r \in R \Rightarrow r \in V$ – элемент может быть представлен в виде правила.
- $\forall p \in P \Rightarrow p = r_1 \circ ... \circ r_n \wedge r_i \in p$ – образец представим сцеплением элементов.

Данные следствия необходимы, чтобы ввести единую для элементов, образцов и правил извлечения функцию покрытия $a : T \times V \to \{истина, ложь\}$. Функция покрытия $a(t, v)$ для любого правила $v \in V$ и любого текстового сегмента $t \in T$ позволяет ответить на вопрос, покрывает ли правило данный сегмент.

Чтобы дать интерпретацию функции покрытия для элементов образцов, рассмотрим структуру элемента: $r_i =< c, e, l_1, l_2 >$, где $c \subseteq W$ – лексическое ограничение элемента, $e \subset W$ – исключение лексического ограничения, l_1 и l_2 – минимальная и максимальная длина покрытия элемента.

Лексическое ограничение c и его исключение e определяют множество слов $c \backslash e \subset W$, которые могут встречаться в текстовых сегментах $T_{ri} \subset T$, покрываемых элементом r_i. Значения l_1 и l_2 определяют допустимый диапазон длин текстовых сегментов T_{ri}. Функция покрытия $a(t, r)$ для элементов и $a(t, p)$ для образцов имеет вид.

$$a(t,r) = true \Leftrightarrow \begin{cases} r =< c, e, l_1, l_2 > \\ (\forall w \in t \Rightarrow w \in c \wedge w \notin e) \vee t = t_\varnothing \\ l_1 \le N_t \le l_2 \end{cases} \qquad a(t,p) \Leftrightarrow \begin{cases} p = r_1 \circ r_2 \circ ... \circ r_n \\ t = t_1 \bullet t_2 \bullet ... \bullet t_n \quad \vee (t = t_\varnothing \wedge p = p_\varnothing) \\ \forall i = 1..n \Rightarrow a(t_i, r_i) \end{cases}$$

Образец $p = r_1 \circ r_2 \circ ... \circ r_n$ покрывает сегмент $t \in T$, если $t \in T$ может быть представлен в виде сцепления $t = t_1 \bullet t_2 \bullet ... \bullet t_n$, так что t_i-ый сегмент покрывается соответствующим r_i-ым элементом. При этом допускается, что некоторые сегменты из $t_1, t_2, ..., t_n$ могут быть пустыми. Данный факт позволяет судить об образцах, как о шаблонах фраз, состоящих из элементов, связанных отношением предшествования. Функция покрытия для правила извлечения задается следующим образом:

$$a(t,v) \Leftrightarrow \begin{cases} v = p_b \circ p_c \circ p_a \wedge t = t_b \bullet t_c \bullet t_a \\ a(t_b, p_b) \wedge a(t_c, p_c) \wedge a(t_a, p_a) \\ t_c - результат\ извлечения \end{cases}$$

Т.е. правило покрывает текстовый сегмент $t \in T$, если он представим в виде сцепления трех сегментов $t_b \bullet t_c \bullet t_a$, каждый из которых покрывается соответствующим образцом правила. Извлечению подлежит только часть t_c, которая объявляется значением фреймового слота. Между правилами извлечения и схемой фреймов устанавливается следующая связь:

– $\forall s\ \exists V_s \subset V : \forall v \in V_s \wedge \forall t \in T \wedge a(t,v) = истина \Rightarrow t \in T_i : s R_{ST} T_i$,

– $\forall s_1, s_2\ \exists V_{s1}, V_{s2} \subset V : V_{s1} \cap V_{s2} = \varnothing$.

В главе также описаны три способа описания лексических ограничений и их исключений: в виде множества непосредственно перечисленных слов, в виде объединения орфографических признаков и в виде пересечения морфологических признаков. Любой способ описания должен удовлетворять требованию – множество C лексических ограничений должно образовывать алгебраическую решетку CL. Данное требование необходимо, чтобы существовала возможность обучения модели EM, этот факт сформулирован и доказан в виде теоремы «О существовании и поиске модели извлечения».

В качестве реализации модели предложено использовать недетерминированный автомат без циклов, для которого приведен алгоритм формирования топологии на основе модели EM. Также разработан алгоритм извлечения, использующий метод ветвей и границ для выполнения поиска ограниченного числа правил извлечения, покрывающих анализируемый сегмент.

<u>В главе 5</u> «Разработка метода обучения модели извлечения» приводится описание метода формирования правил извлечения на основе обучающих примеров. Метод следует стратегии А3 (см. рис. 2) и реализует второй этап работы системы извлечения (см. рис. 1). В качестве вербализатора выступает человек-эксперт, выполняющий разметку текстов, формируя тем самым обучающую выборку. Метод включает четыре этапа: формирование предельно конкретных правил, групповое сжимающее обобщение, унарное обобщение и формирование исключений элементов правил.

Все этапы объединяет требование о возрастании функции качества $F(V,T_e) = \frac{1}{N_v} \sum_{v \in V} f(v,T_e)$, где $f(v,T_e) = \frac{2 \cdot a(v,T_e)}{b(v,T_e) + N_p}$ – F-мера качества извлечения правила v, $b(v,T_e)$ – общее количество извлеченных сегментов правилом v, из которых $a(v,T_e)$ – число корректных извлечений, N_p – количество позитивных обучающих примеров, которые должно покрыть в идеале правило v.

На первом этапе из каждого примера формируется правило извлечения, покрывающее только этот пример. На втором этапе эти предельно конкретные правила подлежат обобщению. Второй этап обучения реализует групповую сжимающую стратегию следующим образом. На каждом шаге обучения в текущем наборе правил выделяются группы, заменяемые группами обобщенных правил. Это достигается за счет построения графа $G = (V_m, V_g, R_{mg})$, так что его вершинам соответствуют правила текущего набора V_m, а с каждым ребром связано правило $v_{ij} \in V_g$, полученное в результате

парного обобщения правил $v_i \in V_m$ и $v_j \in V_m$, вершины которых соединяет данное ребро. Граф задан матрицей смежности так, что $G[v_i, v_j] = G[v_j, v_i] = v_{ij}$.

Полученный граф анализируется на предмет наличия циклов. Для каждой вершины v_i графа находится оптимальный цикл $C_i = v_i ... v_k ... v_i$ такой, чтобы для каждой вершины v_k цикла из всех ребер $v_{ks} = G[v_k, v_s]$ графа, инцидентных ей, циклу принадлежало бы ребро $v_{kl} = \underset{\forall v_{ks} = G[v_k, v_s]}{arg\,max}\ \beta(v_{ks}, T_e)$.

Комплексный показатель качества правила $\beta(v_{kl}, T_e)$ рассчитывается как $\beta(v_{kl}, T_e) = K_f \cdot f(v_{kl}, T_e) + K_\rho \cdot \rho(v_{kl}) + K_S \cdot S(v_{kl})$, где $f(v_{kl}, T_e)$ – F-мера извлечения правила v_{kl}; $\rho(v_{kl})$ – качество обобщения правила v_{kl}; $S(v_{kl})$ – абсолютная степень специфичности правила. Мера $\rho(v_{kl})$ количественно оценивает погрешность обобщения пары правил v_k и v_l при формировании нового правила v_{kl}. Мера $S(v_{kl})$ учитывает количество элементов в правиле и мощности их лексических ограничений так, что чем больше информации заложено в правиле, тем больше значение $S(v_{kl})$. Числовые коэффициенты K_f, K_ρ и K_S определяют значимость соответствующей составляющей комплексного показателя $\beta(v_{kl}, T_e)$. При проведении экспериментов коэффициенты выбирались в следующих отношениях $K_f > K_\rho > K_S$.

Правила, соответствующие вершинам найденного цикла, заменяются обобщенными правилами, представляющими его ребра. С каждой такой заменой выполняется как сокращение набора правил извлечения, так и увеличение F-меры отдельных правил, так что возрастает мера $F(V, T_e)$.

Особенностями предложенного алгоритма являются следующие:

– размер обобщаемой группы правил не фиксирован и определяется длинами оптимальных циклов, которые для каждой вершины графа являются различными;
– группа обобщенных правил, соответствующих вершинам цикла, заменяется не на одно правило, а на целую группу правил, соответствующих ребрам цикла.

В экспериментальной части диссертации показано, что описанное усовершенствование сжимающей стратегии придает обученной модели уникальные свойства, одним из которых является малое отставание полноты извлечения от точности, что свидетельствует о возможности эффективно увеличивать качество извлечения за счет наращивания числа обучающих примеров.

Тем не менее, в основе группового обобщения лежат парные обобщения правил v_k и v_l, что сведено к независимому обобщению их образцов. Более точно задача обобщения образцов формулируется так: имеются два образца $p_i = r_1 \circ r_2 \circ ... \circ r_N$ и $p_j = q_1 \circ q_2 \circ ... \circ q_M$, необходимо на их основе сформировать множество обобщенных образцов P_{ij}, обеспечив при этом минимальную погрешность обобщения. Для решения этой задачи введена матрица соответствий A размерностью $N \times M$, где строкам соответствуют элементы образца $p_i = r_1 \circ r_2 \circ ... \circ r_N$, а столбцам – элементы образца $p_j = q_1 \circ q_2 \circ ... \circ q_M$, при

этом $A[i,j]=c_{ij}$, где $c_{ij}=c_i \veebar c_j$ – лексическое ограничение, полученное операцией наименьшей верхней границы $\veebar$ решетки лексических ограничений CL, примененной к лексическим ограничениям элементов $r_i =< c_i, \varnothing, l_1^i, l_2^i >$ и $q_j =< c_j, \varnothing, l_1^j, l_2^j >$ исходных образцов. С каждой ячейкой также связано значение $s(c_{ij})=1-err(c_{ij})$, где $err(c_{ij})=\sum_{w\in c_i \veebar c_j} p(w) - \sum_{w\in c_i \cup c_j} p(w)$ – погрешность обобщения $c_{ij}=c_i \veebar c_j$ по решетке CL, $p(w)$ – вероятность встретить слово w в текстах предметной области. Далее формирование обобщенных образцов на основе p_i и p_j сводится к поиску маршрутов $P=a_1,a_2,...,a_L$ опорных ячеек матрицы, удовлетворяющих требованиям:

– $\forall a_i, a_{i+1} \in P : a_i = A[i_1, j_1] \wedge a_{i+1} = A[i_2, j_2] \Rightarrow (i_1 < i_2 \wedge j_1 < j_2)$,

– $\forall a_i \in P : a_i = A[k,l] \Rightarrow s(c_{kl}) > 0$.

Анализ матрицы сводится к поиску опорных маршрутов длины $1 \le L \le min(N,M)$ наиболее близких в геометрическом смысле к эталонному маршруту той же длины.

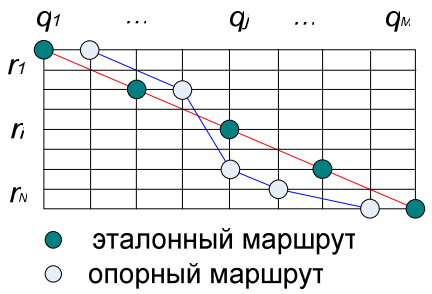

Рис. 3. Маршруты матрицы *A*

Эталонный маршрут размещен на диагонали матрицы так, что его ячейки равноудалены друг от друга (см. рис. 3). Близость маршрутов P_s и P_e рассчитывается по формуле:

$$\rho(P_s,P_e)=\sqrt{\sum_{l=1}^{L} s\left(A[i_l^s; j_l^s]\right)^2 \cdot H\left(i_l^s, i_l^e\right) \cdot H\left(j_l^s, j_l^e\right)},$$ где

$$H(i_1,i_2)=-\frac{i_1}{i_1+i_2}\cdot log_2\left(\frac{i_1}{i_1+i_2}\right)-\frac{i_2}{i_1+i_2}\cdot log_2\left(\frac{i_2}{i_1+i_2}\right)$$ – функция, подобная энтропии, позволяющая оценить геометрическую близость для произвольной пары ячеек матрицы, чем сильнее отличаются индексы i_1 и i_2, тем ближе значение $H(i_1,i_2)$ к нулю. Элементы обобщенного образца формируются на основе ячеек найденного опорного маршрута двумя правилами $G_1(r_k;q_l)$ и $G_2\begin{pmatrix} r_{n+1}\circ...\circ r_{k-1} \\ q_{m+1}\circ...\circ q_{l-1}\end{pmatrix}$.

– $G_1 \equiv (d =< c, \varnothing, l_1, l_2 >) : c = A[k,l] \wedge l_1 = min\left(l_1^k, l_1^l\right) \wedge l_2 = max\left(l_2^k, l_2^l\right)$.

– $G_2 \equiv (d =< c, \varnothing, l_1, l_2 >) : c = \underset{r=n+1}{\overset{k-1}{\veebar}} c_r \veebar \underset{q=m+1}{\overset{l-1}{\veebar}} c_q \wedge l_1 = min\left(\sum_{r=n+1}^{k-1} l_1^r, \sum_{q=m+1}^{l-1} l_1^q\right) \wedge l_2 = max\left(\sum_{r=n+1}^{k-1} l_2^r, \sum_{q=m+1}^{l-1} l_2^q\right)$.

Правило G_1 формирует элемент нового образца на основе элементов $r_k \in p_i$ и $q_l \in p_j$ исходных образцов, соответствующих опорным ячейкам матрицы. Правило G_2 формирует элемент нового образца на основе фрагментов исходных образцов $r_{n+1}\circ...\circ r_{k-1} \subset p_i$ и $q_{m+1}\circ...\circ q_{l-1} \subset p_j$, размещенных между парами элементов, соответствующих двум соседним опорным ячейкам маршрута.

Глава 6 «Разработка модели морфологического анализа и метода ее обучения» посвящена описанию модифицированного принципа аналогии и разработанной на его основе обучаемой модели морфологического анализа.

Данная разработка обусловлена необходимостью качественного морфологического анализа текстов на третьей стадии работы системы извлечения (см. рис. 1).

Оригинальный принцип аналогии, предложенный Г.Г. Белоноговым и А.А. Хорошиловым, формулируется следующим образом: «Слова с одинаковыми концевыми буквосочетаниями имеют одинаковые морфологические признаки». Предложенная модификация действует согласно парадигме «Слова с одинаковыми концевыми буквосочетаниями имеют одинаковые морфологические признаки, а также одинаковое концевое буквосочетание канонических форм».

Модификация позволяет избавиться от необходимости поддерживать емкий словарь словоизменительных основ для обеспечения синтеза канонической формы. Предложенный метод отличается следующими положениями.

- Для определения канонической формы слова используется словарь концевых буквосочетаний канонических форм (образцы канонических форм), который существенно меньше исчерпывающего словаря словоизменительных основ.
- При синтезе канонической формы у слова отсекается характерное для его морфологических признаков концевое буквосочетание (образец слова) и присоединяется образец канонической формы, при этом существенно повышается производительность анализатора.

Предложенная модель морфологического анализатора имеет вид $MA = < P_w, P_c, T_m, R_{wcm}, \alpha >$, где P_w и T_m – множества образцов слов и морфологических шаблонов (наборов морфологических признаков слов), P_c – множество образцов канонических форм, R_{wcm} – отношение на $P_w \times P_c \times T$, α – отображение вида $\alpha : W \Rightarrow R_{wcm} : \forall w \in W \; \exists \alpha(w) \subset R_{wcm}$. Отношение R_{wcm} позволяет задать связь образцов слов и канонических форм с морфологическими шаблонами так, что каждая тройка $(p_w, p_c, t) \in R_{wcm}$ отражает тот факт, что у любого слова, соответствующего образцу p_w и имеющего морфологический шаблон t, каноническая форма соответствует образцу p_c.

Морфологический анализ слова w сводится к следующим действиям:

- определение тройки $(p_w, p_c, t) \in \alpha(w)$ для слова w такой, что p_w является максимально длинным образцом слова w;
- замена p_w в слове w на p_c с образованием канонической формы c.

В проводимых экспериментах словарь обученного морфологического анализатора содержал ~130 тыс. образцов, что на порядок меньше совокупного объема словаря, использованного в работе Белоногова и Хорошилова.

В главе также описан метод обучения модели MA. Метод полностью отвечает стратегии А3 (см. рис. 2) и реализует первый этап работы системы извлечения (см. рис. 1).

Особенностями предложенного метода обучения являются следующие.

1. В качестве вербализатора выступает не человек, а другой морфологический анализатор-учитель. Таким образом, человек лишен необходимости вручную формировать обучающие примеры.

2. Анализатор-учитель должен обладать высокой точностью, однако к его полноте жестких требований не предъявляется.

Особенностью получаемого в результате обучения морфологического анализатора является высокая F-мера качества его работы в сравнении с качеством анализатора-учителя, т.к. обученный анализатор имеет как высокую точность, так и высокую полноту.

Метод обучения основывается на большом массиве слов, корректно разобранных учителем (см. рис. 4). В экспериментах этот массив содержал ~430 тыс. слов, из которых было сформировано ~830 тыс. обучающих примеров. Обучение выполняется в три этапа.

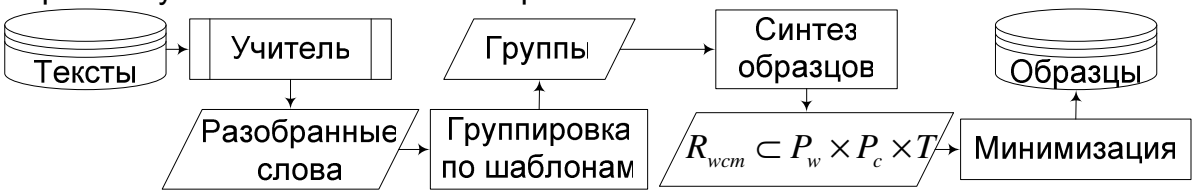

Рис. 4. Этапы метода обучения модели MA

На первом этапе примеры группируются по морфологическим шаблонам. На втором этапе в каждой группе выделяются образцы, объединяющие примеры с совпадающими концевыми буквосочетаниями на основе четырех критериев:

1. $\forall w_1, w_2 \in W_t : \exists p_{w12} \Rightarrow max(l_{pw12})$ – Максимизация длины общего для пары слов (w_1, w_2) концевого буквосочетания.
2. $\forall w_1, w_2 \in W_t : \exists p_{w12} \Rightarrow \exists p_{c1}, p_{c2} : p_{c1} = p_{c2}$ – Образцы канонических форм слов w_1 и w_2 должны совпадать.
3. $\forall w_1, w_2 \in W_t : \exists p_{w12} \wedge \exists \alpha(w_1) \neq \varnothing \Rightarrow \left(\forall (p_{wi}, p_{ci}, t_i) \in \alpha(w_1) \ : w_1 \notin W_{ti} \Rightarrow l_{pwi} < l_{pw12}\right)$ – Прежде чем генерировать образец p_{w12} на основе слова w_1, необходимо проверить, не удовлетворяет ли w_1 существующим образцам в других группах W_{ti}. Если для w_1 удается найти существующие образцы $\{p_{wi}\}$, то необходимо потребовать, чтобы p_{w12} был длиннее всех $\{p_{wi}\}$.
4. $\forall w_1, w_2 \in W_t : \exists p_{w12} \wedge \exists w \in W_{ti} \neq W_t : (p_{w12}, p_{c12}, t) \in \alpha(w) \Rightarrow \exists (p_w, p_c, t) \in \alpha(w) : l_{pw} > l_{pw12}$ – После генерации образца p_{w12} на основе слова w_1 необходимо проверить, не существует ли среди уже проанализированных слов, относящихся к другим морфологическим шаблонам W_{ti}, такого слова w, которое также удовлетворяло бы образцу p_{w12}. Если такое слово найдено, то необходимо убедиться, что его образец p_w длиннее, чем p_{w12}, в противном случае необходимо отказаться от принятия p_{w12} и пытаться создать более короткий образец на основе w_1.

На последнем этапе обучения выполняется постепенное сокращение максимизированных на предыдущем этапе длин образцов.

В главе 7 «Экспериментальное исследование свойств разработанных моделей и методов» приведены результаты экспериментов, где исследованы свойства обучаемой модели морфологического анализа MA и свойства обучаемой модели извлечения EM.

Для модели *MA* проанализирована зависимость объема и структуры словаря образцов от объема обучающей выборки. Показано, что с ростом обучающей выборки кривая объема итогового словаря имеет точку перегиба, начиная с которой с ростом числа примеров количество образцов убывает. Причиной такого эффекта являются два фактора: рост числа парных обобщений по сравнению с унарными и, как следствие, более эффективная минимизация словаря.

Основная масса образцов слов распределена в диапазоне длин от 4 до 11 символов. На формирование этих образцов приходится 90% всех обучающих примеров. Точность и полнота анализатора на известных словах зависит линейно от объема обучающей выборки и при максимальной выборке достигают абсолютных значений, близких к 1. Обученный анализатор в состоянии разобрать 70% слов, неизвестных учителю.

Эксперименты с обучением модели извлечения позволили сделать следующие выводы. Для текстов рассмотренных жанров наблюдается сравнительно малое отставание полноты извлечения от точности, что является уникальным в сравнении с зарубежными аналогами. В зависимости от жанра текстов характер зависимости показателей качества от объема обучающей выборки различен. Кроме того, для достижения значения F-мерой отметки 0,85 требуется разное количество примеров. Чем меньше степень свободы автора при составлении текста, тем сильнее ограничена форма контекстов, окружающих извлекаемые значения фреймовых слотов. Как следствие, требуется меньшее количество примеров для формирования правил извлечения, покрывающих тестовую выборку. Так для текстов телеграммного жанра потребовалось всего 40 примеров, чтобы добиться значения F=0,98, для текстов жанра официальных документов F=0,85 после обучения на 250 примерах, тогда как для жанра информационной заметки F=0,85 достигается после обучения на 1000 примерах.

В целом результаты экспериментов показали адекватность разработанных моделей и практическую применимость методов их обучения для задачи извлечения знаний из текстов.

В заключении приводятся выводы по работе в целом и формулируются основные результаты, полученные в диссертации.

В приложениях приведены описания двух систем сопоставляющего анализа, реализованных на основе предложенных в диссертации моделей и методов. Приведены их функциональные схемы и примеры результатов работы на реальных данных.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Достигнута поставленная цель работы – разработаны модели и методы извлечения знаний из текстов, применимые как для задач сопоставляющего анализа, так и для других задач интеллектуальной обработки текстов. Основными результатами работы являются следующие.

1. Предложена комбинированная модель представления знаний, сочетающая современные достижения в области онтологического представления

знаний и достоинства фреймовых моделей и предоставляющая инструмент для разработки систем сопоставляющего анализа.

2. Разработана модель извлечения знаний из текстов и метод ее обучения. В модели выделена решетка лексических ограничений, на основе которой доказана теорема о возможности обучения модели извлечения. Простота структуры правил извлечения обеспечивает практическую реализуемость механизмов обучения, а также обеспечивает реализацию метода извлечения на основе конечного автомата.
3. Предложена модель морфологического анализа и метод ее обучения. В качестве учителя используется другой морфологический анализатор с высокими показателями качества на неполном лексиконе языка. Обучение не требует вмешательства человека. Важное свойство обученной модели – способность разбирать изначально неизвестные слова.
4. Все разработанные модели и методы доведены до программной реализации в виде самостоятельных продуктов для использования в задачах автоматизированной обработки текстов.
5. Разработан программный комплекс для выполнения экспериментальной проверки работоспособности моделей. Проведенные эксперименты показали: для морфологического анализа точность обученной модели составляет 0,99. Для модели извлечения значение 0,85 F-меры качества достигается на 30% обучающих примеров от общего числа примеров для текстов жанра информационной заметки. Для текстов телеграммного жанра F-мера достигает значения 0,98 на 20% обучающих примеров.
6. Разработанные модели и методы:
   6.1. Внедрены в Системе семантического контроля текстов редактируемых документов, используемой в Совете Федерации Федерального Собрания Российской Федерации для выявления кадровых несоответствий в текстах стенограмм заседаний Совета Федерации;
   6.2. Внедрены в Информационно-поисковой системе «Обзор СМИ» в части автоматического построения аннотаций к документам, используемой в Совете Федерации Федерального Собрания Российской Федерации;
   6.3. Используются в Интеллектуальной системе выявления и исправления ошибок в почтовых адресах, разработанной компанией НПЦ «ИНТЕЛТЕК ПЛЮС»;
   6.4. Использованы в НИР по кластеризации и классификации текстовых документов для систем специального назначения, выполненной ВНИИНС для МО.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Симаков К.В. Метод обучения модели извлечения знаний из естественно-языковых текстов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Вестник МГТУ. Приборостроение.–2007. – №3.– С. 75–94.
2. Симаков К.В. Модель извлечения знаний из естественно-языковых текстов / А.М. Андреев, Д.В. Березкин, К.В. Симаков // Информационные технологии. – 2007. – №12. – С. 57–63.

3. Андреев А.М., Березкин Д.В., Симаков К.В. Архитектура системы машинного понимания текстов // Информатика и системы управления в XXI веке: Сборник трудов – М.: Изд-во МГТУ им. Н.Э. Баумана, 2003. – №1. – С. 419–423.
4. Березкин Д.В., Симаков К.В. Формальный V - язык описания морфологии и синтаксиса текстов на естественном языке // Информатика и системы управления в XXI веке: Сборник трудов – М.: Изд-во МГТУ им. Н.Э. Баумана, 2003. – №1. – С. 364–368.
5. Андреев А.М., Березкин Д.В., Симаков К.В. Снятие синтаксической омонимии в задачах машинного понимания естественных текстов // Информатика и системы управления в XXI веке: Сборник трудов – М.: Изд-во МГТУ им. Н.Э. Баумана, 2003 – №1. – С. 415–418.
6. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа / А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой всероссийской научной конференции (RCDL'2003) – Санкт-Петербург: НИИ Химии СПбГУ, 2003. – С. 140–149.
7. Андреев А.М., Березкин Д.В., Симаков К.В. Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды шестой всероссийской научной конференции (RCDL'2004) – Пущино, 2004. – С. 93–102.
8. Андреев А.М., Березкин Д.В., Симаков К.В. Обучение морфологического анализатора на большой электронной коллекции текстовых документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды седьмой всероссийской научной конференции – Ярославль: Ярославский государственный университет, 2005. – С.173–181.
9. Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды восьмой всероссийской научной конференции (RCDL'2006) – Ярославль: Ярославский государственный университет, 2006. – С. 252–262.
10. Использование технологии Semantic Web в системе поиска несоответствий в текстах документов / А.М. Андреев, Д.В. Березкин, В.С. Рымарь, К.В. Симаков // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды восьмой всероссийской научной конференции (RCDL'2006) – Ярославль: Ярославский государственный университет им. П.Г. Демидова, 2006. – С. 263–269.
11. Автоматизация обнаружения и исправления опечаток в названиях географических объектов для системы семантического контроля документов электронной библиотеки / А.М. Андреев, Д.В. Березкин, А.С. Нечкин, К.В. Симаков, Ю.Л. Шаров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды девятой всероссийской научной конференции (RCDL'2007) – Переславль-Залесский: Университет города Переславль, 2007. – Т.2. – С. 49–56.