На правах рукописи

УДК 519.68; 681.513.7;
612.8.001.57; 007.51/.52

БАТУРА
Татьяна Викторовна

МАШИННО-ОРИЕНТИРОВАННЫЕ ЛОГИЧЕСКИЕ МЕТОДЫ ПРЕДСТАВЛЕНИЯ СМЫСЛА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико – математических наук

Новосибирск 2006

Работа выполнена в Институте систем информатики
имени А.П. Ершова СО РАН

Научный руководитель: Мурзин Федор Александрович,
кандидат физико – математических наук

Официальные оппоненты: Непейвода Николай Николаевич,
доктор физико – математических наук

Викентьев Александр Александрович,
кандидат физико – математических наук

Ведущая организация: Сибирский государственный университет телекоммуникаций и информатики

Защита состоится 23 июня 2006 г. в 15 ч. 30 мин. на заседании диссертационного совета К.003.032.01 в Институте систем информатики имени А.П. Ершова Сибирского отделения РАН по адресу: 630090, г. Новосибирск, пр. ак. Лаврентьева, 6.

С диссертацией можно ознакомиться в читальном зале ИСИ СО РАН (г. Новосибирск, пр. ак. Лаврентьева, 6).

Автореферат разослан _____ 2006 г.

Ученый секретарь
Диссертационного совета,
к.ф.–м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Исследования в области автоматической обработки текста и формализации естественных языков, планомерно продвигаясь от самих простых методов анализа к более сложным, постепенно приближаются к такому уровню обработки текста, на котором уже возможно представление текста не просто в виде последовательности слов, а единым целым, обладающим неким смыслом, что уже соответствует человеческому восприятию.

В настоящее время происходит активизация в области исследования лингвистических проблем формальными методами и в области применения для этих целей компьютеров. Это, прежде всего, связано с ростом производительности вычислительных систем, что позволяет в реальное время выполнять алгоритмы обработки текстов, которые раньше выполнить в реальное время было невозможно.

С лингвистической точки зрения представляются актуальными различные модели смысла текста, и более широко – различные подходы к отображению семантики текстов на естественном языке. Поэтому была предпринята попытка исследовать смысл текста, исходя из предварительного структурного разбора этого текста.

В диссертации речь идет о методах, которые позволят проводить разносторонний анализ текстов и отдельных предложений на естественном языке. Рассматриваются такие методы, как представление смысла текста в рамках подхода И.А. Мельчука и предложенные им лексические функции, теоретико-множественные модели С. Маркуса; также делается попытка адаптировать для целей изучения текстов на естественном языке некоторые методы и конструкции математической логики: конструкцию Хенкина, применяемую в теореме о существовании модели и в теоремах об опускании типов и др. В работе представлены разнообразные алгоритмы сопоставления предикатов и формул узкого исчисления предикатов текстам на естественном языке. Кроме этого, некоторые конечные модели сопоставлены предложениям текста и тексту целиком.

Представляется актуальной задача разработки методов, которые бы позволили использовать при машинной обработке толковые словари, обычные «человеческие», а не машинно-ориентированные. В работе предпринята попытка анализа структуры словарных статей словаря С.И. Ожегова.

Актуальной является задача исследования процесса освоения человеком речи на ранних этапах его развития с целью моделирования этого процесса на компьютере, проблема понимания, с какими структурами данных целесообразно иметь дело при обработке лингвистической информации, и как подобные структуры и алгоритмы работы с ними могут быть поддержаны аппаратно, в том числе с использованием параллелизма и др. В диссертации уделено много внимания этим вопросам.

Теория языка как структуры, соответствующая классификация и методы обработки формальных языков начали разрабатываться в математике (а позднее и в информатике), еще с 30-х годов. Однако прямое применение существующего аппарата описания формальных языков к естественному языку невозможно, из-за того что это объект принципиально другой природы. В частности, в отличие от формального языка, естественный язык не следует задуманной и последовательно реализованной концепции. Он развивается с течением времени под воздействием многих внешних и внутренних сил, становится тем, что он из себя представляет, и усваивается в сообществе через использование в коммуникации, а не благодаря правилам. Кроме того, чисто грамматическое описание естественного языка не достаточно для использования, поскольку естественный язык не является просто вещью в себе, он необходимо соотносится со структурами знания, используемыми его носителями. В результате описание грамматики естественного языка как некоторого класса формальной грамматики оказывается затруднено, что все же не отменяет полезность классификации формальных грамматик для компьютерной лингвистики.

С другой стороны, для того чтобы допускать возможность реальной компьютерной реализации, лингвистическая теория должна обладать высокой степенью формализации и полноты. Поэтому общей чертой для всех теорий, используемых в компьютерной лингвистике, является их ге-

нативность в том смысле, что исследование естественного языка ведется через построение полностью явных описаний и определение общей структуры пространства этого описания. Кроме того, реализация лингвистической теории через инструментальную систему для описания структур естественного языка зависит также от методов программирования, использованных для ее написания. Таким образом, развитие компьютерной лингвистики стимулируется, с одной стороны, развитием теоретических средств описания естественного языка, а с другой – прогрессом технологий программирования, в первую очередь, в области искусственного интеллекта.

Если понятие инструментального средства рассмотреть в контексте классического различия, проводимого в лингвистике между языковой компетенцией и использованием языка его носителем, то можно отметить, что, во-первых, это инструментальное средство должно обладать возможностью представлять знание о языке, во-вторых, в нем должно быть организовано использование этого знания, для того чтобы понимать и/или генерировать конкретный текст на естественном языке. Иными словами, идеальная инструментальная система обработки естественного языка должна основываться на идеальной лингвистической теории, т. е. обладать средствами представления лингвистических структур, структур представления знаний, а также на алгоритмах для обработки таких структур. Она, в частности, должна поддерживать возможность представления сложных средств выражения, свойственных естественному языку, таких как лексические омонимия и полисемия (несколько значений, соответствующих одному слову), синонимия (несколько слов имеют близко связанные значения), привязка к контексту речи (с помощью анафорических местоимений) и к контексту ситуации (экзофорические или дейктические указатели), эллипсис (как синтаксический, так и семантический), фигуры речи (использование слов не в их прямом значении) и т. д. Традиционной проблемой является также описание взаимосвязи между грамматическими структурами и содержанием предложения, при этом содержание представлено либо как логическая формула, либо как структура, записанная на некотором языке представления знаний.

Совокупная сложность вышеописанных феноменов существенно выше существующих на данном этапе теоретических построений для их описания, обладающих требуемой степенью полноты и формализации. Иными словами, не существует ни идеальной теории для компьютерной лингвистики, ни идеальных средств ее реализации. По этой причине невозможно создать идеальную инструментальную систему для обработки естественного языка, что приводит к изобилию существующих систем. Чаще всего набор средств представления инструментальной системы (а также полнота этого набора) определяется теоретической моделью, лежащей в ее основе.

Цель работы

Цель работы – разработка методов, позволяющих проводить разносторонний анализ текстов и отдельных предложений на естественном языке, в том числе, позволяющих осуществлять представление смысла текстов и предложений.

Для достижения поставленной цели в данной работе было необходимо решить следующие задачи:

- разработать различные алгоритмы сопоставления предикатов и формул логики первого порядка предложениям на естественном языке;
- рассмотреть возможность сопоставления конечных моделей предложениям текста и тексту целиком;
- рассмотреть возможность применения общих принципов организации памяти с параллельным доступом к обработке лингвистической информации;
- проанализировать структуру словарных статей толкового словаря С.И. Ожегова и рассмотреть возможность представления предложений на естественном языке в виде деревьев с пометками;
- проанализировать процесс формирования речи у человека, выделить основные этапы когнитивного развития и разработать формальные модели базовых конструкций языка.

Методы исследования

В основном, применялись методы, относящиеся к информационным технологиям и используемые при обработке текстов на естественном языке, и методы из математической логики. Также был привлечен довольно обширный материал из классической и математической лингвистики, психологии развития и антропологии.

Научная новизна

Проведенные исследования позволили разработать и предложить ряд новых определений, формальных моделей и алгоритмов, которые могут быть применены при анализе и обработке текстов на естественном языке.

В частности, разработаны алгоритмы сопоставления различных предикатов и формул логики первого порядка предложениям на естественном языке, основанные на использовании грамматической и синтаксической структуры слов и предложений. Предложено использовать конструкцию Хенкина из математической логики для построения конечных моделей, которые могут трактоваться как смысл текста.

На основе проведенного анализа структуры словарных статей толкового словаря С.И. Ожегова предложен механизм представления предложений в виде деревьев с пометками, который может быть использован в поисковых системах. Проанализированы словарные статьи из упомянутого выше словаря, относящиеся к временным конструкциям и понятиям, связанным с местоположением.

Обоснована возможность применения модификаций конструкций языка символьных преобразований REFAL для формирования деревообразного представления предложений на естественном языке и схем «вопрос-ответ» и описан алгоритм использования их в поисковых системах. Приведен большой список, более сорока схем типа «вопрос-ответ», которые могут быть полезны при реализации программных систем, ориентированных на обработку текстов.

Рассмотрены основные этапы формирования речи у человека на ранней стадии развития, и как результат, предложена формализованная модель конструкций языка, называемых базовыми. Проведены предвари-

тельные исследования относительно применения принципов организации памяти с параллельным доступом к обработке лингвистической информации.

Практическая ценность

Результаты работы могут быть применены в автоматизированных системах акцепции информации из текстов на естественном языке, интеллектуальных системах поиска информации в сети, при построении систем автоматического резюмирования, электронных переводчиков и словарей. Предполагается использование результатов работы в системах безопасности, работающих с банковской информацией. К вопросу о размещении данных в памяти с параллельным доступом и некоторым другим вопросам, затронутым в диссертационной работе, проявила интерес корпорация IBM. Она выделила грант на поддержку данной работы.

Апробация работы

Результаты работы были представлены на IV Международной конференции по вычислительным наукам, проходившей в Польше в Кракове; докладывались на конференциях-конкурсах «Технологии Microsoft в информатике и программировании», проходивших в Новосибирске в 2004 – 2006 годах и на международных научных студенческих конференциях «Студент и научно-технический прогресс» в 2003 – 2006 годах, в Институте систем информатики имени А.П. Ершова СО РАН, Институте математики имени С.Л. Соболева СО РАН и Новосибирском государственном университете, а также на встречах с иностранными специалистами: американскими, французскими и корейскими.

По теме диссертации опубликовано 14 работ.

Структура и объем работы

Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы и трех приложений. Объем диссертации – 184 страницы. Список литературы содержит 79 наименований. Работа включает 11 рисунков и 3 таблицы.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы исследований и приводится краткое содержание работы.

Первая глава посвящена описанию общих понятий, лежащих в основе современных средств компьютерного представления лингвистической информации. Для компьютерной лингвистики среди формальных языков и грамматик наиболее важны грамматики конечных автоматов, контекстно-свободные и контекстно-зависимые грамматики.

Одна из широко известных теорий отображения смысла текста на естественном языке принадлежит И.А.Мельчуку. При создании модели «Смысл $\Leftrightarrow$ Текст» он вводит понятие лексической функции. Лексическая функция ставит в соответствие каждой из лексических единиц (слов и словосочетаний) набор альтернативных лексических единиц, находящихся с исходной единицей в соответствующем смысловом соотношении.

Среди задач данной диссертационной работы есть задача о сопоставлении предикатов предложениям на естественном языке. В данной главе рассмотрен алгоритм, основанный на лексических функциях, предложенных Мельчуком. Эти функции можно представить на синтаксическом уровне в виде предикатов следующим образом. Если рассмотреть совокупность словоформ в языке, возникающих при склонениях существительных, спряжениях глаголов и т. д. (т. е. весь словарь), и считать, что x и y – слова или словосочетания из этой совокупности, то получаем предикаты следующего вида: $Syn(x, y)$, x, y – синонимы; $Destr(x, y)$, y – типовое название «агрессивного» действия (x = «оса», y = «жалит») и др.

Другой подход к формализации структуры естественных языков – создание теоретико-множественных моделей языков – принадлежит С. Маркусу. Теоретико-множественные модели языков Маркуса строятся следующим образом. Рассматривается некоторое разбиение словаря естественного языка (он считается конечным множеством) на классы (например, совпадающие с множествами флективных форм слов). С помощью такого разбиения можно дать формальное определение грамматического рода или категории падежа. Кроме этого, Маркус вводит понятие синтаксических типов, которые приблизительно соответствуют традиционным частям речи. Осуществляя операции над синтаксическими типами, стано-

вится возможным определить грамматическую правильность предложения на естественном языке.

В данной главе также сделан обзор уже существующих инструментальных систем для описания структур естественного языка. Приведена классификация этих систем с точки зрения их реализации. Системы анализа текстов состоят из графематического, морфологического, фрагментационного, синтаксического и семантического компонентов. Примером такой системы является система ДИАЛИНГ. Еще одна система, рассматриваемая в этой главе, DSTO Fact Extractor System – система для извлечения информации конкретного вида из произвольных текстовых документов.

Так как одной из поставленных в работе задач является задача о рассмотрении возможности применения в обработке текстов общих принципов организации памяти с параллельным доступом к информации, в первую главу включен обзор систем, использующих данный вид памяти.

Во второй главе предложено несколько алгоритмов сопоставления предикатов и формул узкого исчисления предикатов предложениям на естественном языке.

Один из способов введения предикатов – сопоставление частям речи. Предикаты, полученные таким образом, мы назвали грамматическими.

$Adj_2(x, y)$ – категория числа прилагательного: y = «ед», если x – прилагательное в единственном числе, y = «мн», если x – прилагательное во множественном числе.

$Adj_3(x, y)$ – категория рода прилагательных: y = «мр», если x – прилагательное мужского рода, y = «жр», если x – прилагательное женского рода, y = «ср», если x – прилагательное среднего рода.

$(\forall x)\left(Adj_2(x, ед) \leftrightarrow \left(Adj_3(x, мр) \vee Adj_3(x, жр) \vee Adj_3(x, ср)\right)\right)$ – формула обозначает, что если прилагательное в единственном числе, то оно обязательно либо мужского, либо женского, либо среднего рода, и наоборот.

Кроме этого, предикаты можно ассоциировать с членами предложения. Такие предикаты мы назвали синтаксическими.

Одноместные предикаты членов предложения: $P_{sub}(x)$, где x – подлежащее; $P_{obj}(x)$, где x – дополнение; $P_{adv}(x)$, где x – обстоятельство и др. Двухместные предикаты членов предложения: $P_{pred}(x,y)$, x – сказуемое; $P_{attr}(x,y)$, x – определение; y играет роль определяемого слова или словосочетания.

Можно записать формульное представление этих предикатов, считая, что x, y – слова или словосочетания. Верхний индекс в скобках при Q – местность предиката, нижний индекс Q является показателем, от какого члена предложения задается вопрос. Например, если определяемое слово является сказуемым, то от сказуемого можно задать вопрос к обстоятельству $(\forall x,y)\left(Q_2^{(2)}(x,y) \leftrightarrow \left(P_{adv}(y,x) \,\&\, P_{pred}(x) \,\&\, P_{adv}(y)\right)\right)$.

В общем виде формулы для n неоднородных членов предложения записываются следующим образом. Например, формула $(\forall x, y_1, ..., y_n)\left(Q_1^{(n+1)}(x, y_1, ..., y_n) \leftrightarrow \left(\&_{i=1}^{n} P_{attr}(y_i, x) \,\&\, P_{sub}(x) \,\&\, \&_{i=1}^{n} P_{attr}(y_i)\right)\right)$ означает, что есть неоднородные определения при подлежащем.

В качестве промежуточного результата мы получили, что с помощью введенных предикатов можно определять синтаксические валентности слова. Последние играют немаловажную роль в подходе Мельчука при построении модели «Смысл $\Leftrightarrow$ Текст».

На данном этапе в полученных формулах недостаточно отражена семантическая структура текста, и позже введенные предикаты будут подвергнуты различным преобразованиям средствами математической логики.

В третьей главе описаны структуры данных и потоки, которые удобны для представления предикатов и конечных моделей, сопоставляемых предложениям на естественном языке. Они легко могут быть реализованы средствами языка C++.

Предложению сопоставляем набор структур, состоящих из кортежей, которые в конечном итоге определяют набор предикатов.

С другой стороны, можно считать элементы словаря естественного языка константами, ввести предикаты любым из описанных ранее способом, на основе их получить формулы. Предикаты, в свою очередь, снача-

ла рассматриваем на синтаксическом уровне. Затем смотрим на них уже как на подмножества основных множеств моделей в соответствующих декартовых степенях. Такой подход дает возможность сконструировать модели, т. е. осуществить переход с синтаксического на семантический уровень.

Первоначально предикаты рассматриваем на синтаксическом уровне, т. е. как записи. В дальнейшем на основе полученных структур будут конструироваться модели, т. е. будет осуществлен переход на семантический уровень. Под предикатами в этом случае будем понимать подмножества в соответствующих декартовых степенях основных множеств моделей.

Далее считаем, что на вход поступает текст, т. е. набор предложений. На выходе формируется несколько потоков. В потоки можно записывать информацию о словообразовании. С лексическими функциями тоже могут быть ассоциированы потоки. Конечные модели, сопоставленные исходному тексту, также будем формировать в виде потоков.

Например, все существительные из предложений записываем в поток: $<1, n_1^1, ..., n_{l_1}^1; 2, n_1^2, ..., n_{l_2}^2; ...>$, где последовательно записываются номера предложений и списки существительных, входящих в данное предложение (l_i – длина списка). Причем номер предложения, в котором нет существительных, может быть пропущен.

Перепишем этот поток в другом виде:

$<<1, n_1^1>, ..., <1, n_{l_1}^1>, <2, n_1^2>, ..., <2, n_{l_2}^2>, ...>$.

Обозначим, $C = \left\{ < t, n_j^t > | \, t = \overline{1, N}, j = \overline{1, l_t} \right\}$ – множество всех пар, встречающихся в потоке. Основными множествами моделей будут множества вида $C_0 / \sim$, где $C_0 \subseteq C$, $\sim$ – некоторое отношение эквивалентности. Отношения эквивалентности будут возникать примерно так же, как в конструкции Хенкина при доказательстве теоремы о существовании модели. Пары вида $< t, c_j^t > (t = 1, ..., N)$ могут рассматриваться как константы, и в зависимости от высказываний об этих константах некоторые из них мы объявляем эквивалентными.

В данной главе, кроме того, рассмотрена возможность применения общих принципов организации памяти с параллельным доступом к обработке лингвистической информации.

Рассматриваем предложение на естественном языке как совокупность слов. Слова параллельно подаются в память, т. е. одновременно доступны несколько слов, по одному в каждом модуле. Далее перестановкой строк и столбцов при работе с параллельным доступом к информации можем подбирать нужные сочетания слов.

Другая идея относится к организации структур данных в памяти. Часть слова, состоящая из основы и формообразующего суффикса, запоминается только один раз. В остальных местах, где она встречается, ставится метка. Таким образом, остается хранить только окончания форм слова, что значительно экономнее.

В четвертой главе осуществлен анализ структуры словарных статей из толкового словаря С. И. Ожегова. На основе проведенного анализа предложена специальная деревообразная структура, которая может быть использована для представления любых предложений. Более подробные примеры анализа, а именно, для временных конструкций и для понятий, связанных с местоположением объектов, приведены в конце диссертации в Приложениях 1 и 2.

Деревообразное представление предложений предполагается использовать в поисковых системах следующим образом. Считаем, что поисковый запрос представляет собой совокупность предложений на естественном языке. Эту совокупность предложений можно расширить, используя словарные статьи из толкового словаря (например, словаря Ожегова), т. е. фактически приписать определения отдельных слов. На следующем этапе представляем предложения запроса в виде помеченных деревьев. Вершины помечаем словами, а ребра – вопросами, задаваемыми от одного слова к другому.

При формировании деревообразного представления предложений на естественном языке предлагается использовать конструкции, представляющие собой модификацию конструкций, применяемых в языке символьных преобразований REFAL:

- целесообразно использовать новые типы переменных, связанные с частями речи, частичным совпадением слов и т. д.;
- в языке REFAL для любого оператора $\varphi \to \psi$ выполнено $\mathrm{var}(\varphi) \supseteq \mathrm{var}(\psi)$. У нас это нарушается, и таким образом мы пытаемся учесть контекст.

Далее рассмотрим текст достаточно большого объема, из которого необходимо выбрать предложения по тематике поискового запроса и, таким образом, сформировать аннотацию или решить, является ли текст релевантным данному запросу. Для этого предложения данного текста также могут быть представлены в виде деревьев. После этого необходимо сопоставление на соответствие деревьев из запроса и деревьев, возникших из текста. Для этого в дальнейшем предлагается рассматривать конечные автоматы, работающие на деревьях, по аналогии с подходом Бюхи.

Фактически на данном этапе заданному вопросу соответствует несколько возможных ответов. Поэтому можно считать, что схема перехода «вопрос-ответ» имеет вид $\varphi \to \psi_1 \vee \psi_2 \vee ... \vee \psi_n$.

После отождествления φ с вопросом переменные, входящие в нее приобретают значения. Далее в тексте ищем предложения, которые можно отождествить хотя бы с одной из формул ψ_i. Все такие предложения выдаем пользователю в качестве ответов.

В пятой главе проведен анализ процесса формирования речи у человека. На основе этого анализа произведена периодизация освоения языка ребенком и получены формальные модели базовых конструкций языка.

Под базовыми конструкциями понимаются простейшие в алгоритмическом плане. Они же самыми первыми возникают при освоении речи ребенком. Причем не ставится целью, чтобы эти базовые конструкции «покрыли» весь язык, как, например, базис по типу базиса во множестве булевых функций.

Условно в процессе формирования речи можно выделить пять стадий.

1. На первой стадии формируются отношения эквивалентности, т. е. формируется и запоминается функция вида $f : \omega \to P(\omega)$, где ω – мно-

жество натуральных чисел, $P(\omega)$ – множество конечных подмножеств натуральных чисел.

2. На второй стадии формируются простейшие ассоциации, т. е. формируется функция вида $g:\omega \to P(\omega)$. Но в отличие от первой стадии возникает связь между образами, т. е. между предметами и действиями и их свойствами.

3. На третьей стадии происходит упорядочение ассоциаций $g:\omega \to \langle P(\omega),\leq \rangle$. При этом могут учитываться различные факторы: частота встречаемости данной ассоциации, эмоциональная нагруженность и т. д.

4. На четвертой стадии происходит формирование прототипа грамматики. Имеем функцию $g:\omega \to K$, где K – для каждого слова свое фиксированное множество парадигм этого слова, т. е. K в общем случае состоит из $i_1,...,i_{k_i}$.

Существенным отличием данной стадии от стадии 2 является то, что мы образуем ассоциации, рассматривая близость не в пространстве или во времени, а фактически, работаем с графом, возникающим на стадии 1, т. е. происходит отчуждение слов от образов.

5. На пятой стадии происходит эпизодизация информационного потока. Можно предполагать, что человек не имеет какой-то законченной модели мира, а оперирует эпизодами, как форсирующими условиями в конструкции конечного форсинга, рассматриваемой в математической логике.

Кратко описывается, как для обработки текстов на естественном языке может быть использован конечный форсинг. Считаем, что у нас есть текст, состоящий из предложений на естественном языке. Двигаясь по тексту, получаем истинные и ложные предикаты описанными ранее способами. По некоторым признакам из этих предикатов мы можем формировать множества. Один из таких признаков: если в тексте непротиворечивая информация о чем-либо располагается близко, то речь идет об одном и том же, т. е. синтаксическая близость влечет семантическую. Для противоречивых сведений о чем-то используем другие критерии.

Можно считать, что множество попарно совместных условий образуют верхнюю полурешетку. В ней содержится вся информация из текста. Мы можем рассматривать различные пути, которые будут давать непротиворечивые теории. Таким образом можем получить диаграмму модели, которая, по сути, отражает определенную сюжетную линию в исходном тексте.

В **заключении** перечислены основные результаты работы.

О ЛИЧНОМ ВКЛАДЕ АВТОРА

Исследуемые вопросы являются довольно трудоемкими, а в ряде случаев очень высока их степень неопределенности ввиду отсутствия формализованных постановок. В основном работа выполнялась совместно с научным руководителем. Наибольший вклад автором диссертации внесен в разработку алгоритмов сопоставления различных предикатов и формул логики первого порядка предложениям на естественном языке; в анализ структуры словарных статей толкового словаря С.И. Ожегова; в исследование схем «вопрос-ответ» и описание алгоритма использования их в поисковых системах.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Проведены комплексные теоретические исследования и разработаны методы для проведения анализа текстов и отдельных предложений на естественном языке.

1. Разработаны алгоритмы сопоставления различных предикатов и формул логики первого порядка предложениям на естественном языке и алгоритм сопоставления конечных моделей предложениям текста и тексту целиком, которые могут трактоваться как смысл текста.
2. Проведен анализ структуры словарных статей толкового словаря С.И. Ожегова. На основе проведенного анализа предложен механизм представления предложений в виде деревьев с пометками, который может быть использован в поисковых системах.
3. Предложены конструкции, представляющие собой модификацию конструкций языка символьных преобразований REFAL, которые применимы для формирования деревообразного представления предложе-

ний на естественном языке и схем «вопрос-ответ», и описан алгоритм использования этих схем в поисковых системах.

4. Рассмотрены основные этапы формирования речи у человека на ранней стадии развития, и как результат, предложена формализованная модель конструкций языка, называемых базовыми.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Batura T., Murzin F. Logical Methods for Representing Meaning of Natural Language Texts // Proc. 4th Internat. Conf. on Computational Science – ICCS 2004, Kraków, Poland, June 6-9, 2004. Proceedings, Part III, LNCS 3038. – P. 545 – 551.
2. Батура Т.В., Еркаева О.Н., Мурзин Ф.А. К вопросу об анализе текстов на естественном языке // Новые информационные технологии в науке и образовании. – Новосибирск, 2003. – С.7 – 58.
3. Батура Т. В., Мурзин Ф. А. Логические методы представления смысла текста на естественном языке // Новые информационные технологии в науке и образовании. – Новосибирск, 2003. – С. 59 – 111.
4. Батура Т.В., Корда О.В., Мурзин Ф.А. Исследовательская система для анализа текстов на естественном языке // Методы и инструменты конструирования и оптимизации программ. – Новосибирск, 2005. – С. 7 – 21.
5. Батура Т.В. Представление смысла текста на естественном языке и его лексический анализ // Технологии Microsoft в информатике и программировании. – Новосибирск, 2004. – С. 88 – 90.
6. Батура Т.В. Логический анализ представления смысла текста на естественном языке // Технологии Microsoft в информатике и программировании. – Новосибирск, 2005. – С. 99 – 100.
7. Батура Т.В., Корда О.В., Позименко А.А. Экспериментальная исследовательская система для анализа текстов на естественном языке // Технологии Microsoft в информатике и программировании. – Новосибирск, 2005. – С. 101 – 102.
8. Батура Т.В. Методы логического анализа и представление смысла текста на естественном языке // Технологии Microsoft в теории и практике программирования. – Новосибирск, 2006. – С. 155 – 157.

9. Батура Т.В. Исследовательская система анализа текстов на естественном языке // Технологии Microsoft в теории и практике программирования. – Новосибирск, 2006. – С. 158 – 159.
10. Батура Т.В., Позименко А.А. Система анализа текстов на естественном языке // Материалы XLI междунар. научной студенческой конф. «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск, 2003. – С. 101 – 102.
11. Батура Т.В. Анализ смысла текста на естественном языке // Материалы XLII междунар. научной студенческой конф. «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск, 2004. – С. 195 – 197.
12. Батура Т.В., Корда О.В. Программные средства для анализа текста на естественном языке // Материалы XLII междунар. научной студенческой конф. «Студент и научно-технический прогресс»: Физика. – Новосибирск, 2004. – С. 194 – 195.
13. Батура Т.В. Представление смысла текста на естественном языке с использованием деревообразных структур // Материалы XLIII междунар. научной студенческой конф. «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск, 2005. – С. 92 – 94.
14. Батура Т.В. Применение деревообразного представления текстов в поисковых системах // Материалы XLIV междунар. научной студенческой конф. «Студент и научно-технический прогресс»: Информационные технологии. – Новосибирск, 2006. – С. 159 – 160.

Батура Т.В.

МАШИННО-ОРИЕНТИРОВАННЫЕ ЛОГИЧЕСКИЕ МЕТОДЫ ПРЕДСТАВЛЕНИЯ СМЫСЛА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Автореферат

Подписано в печать Объем 1,1 уч.-изд. л.
Формат бумаги 60 × 90 1/16 Тираж 100 экз.
Отпечатано в ЗАО РИЦ «Прайс-курьер»
630090, г. Новосибирск, пр. ак. Лаврентьева, 6, тел. 34-22-02
Заказ №135