САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Кутарба Анна Юрьевна

ПОСТРОЕНИЕ СЕМАНТИЧЕСКОГО СЛОВАРЯ ДЛЯ ОБРАБОТКИ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

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

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Санкт-Петербург 2006

Работа выполнена на кафедре информатики математико-механического факультета Санкт-Петербургского государственного университета.

Научный руководитель:

доктор физико-математических наук, профессор Тузов Виталий Алексеевич

Официальные оппоненты:

доктор физико-математических наук, профессор Братчиков И.Л.,

кандидат физико-математических наук, доцент Комаров И.И.

Ведущая организация:

Санкт-Петербургский Экономико-математический институт РАН

Защита диссертации состоится “25” мая 2006 года в 14 часов на заседании диссертационного совета Д 212.232.51 по защите диссертаций на соискание учёной степени доктора наук при Санкт-Петербургском государственном университете по адресу: 198504, Санкт-Петербург, Старый Петергоф, Университетский пр., 28, математико-механический факультет, ауд._____

С диссертацией можно ознакомиться в Научной библиотеке имени М.Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб., д. 7/9.

Автореферат разослан “___” апреля 2006 года.

Учёный секретарь
диссертационного совета,
доктор физико-математических наук Мартыненко Б.К.

1. Общая характеристика работы

Актуальность работы. Современное общество часто называют информационным. Действительно, интенсификация обмена информацией идет на самых различных уровнях: от межличностного до межгосударственного. Несмотря на распространение знания иностранных языков, в первую очередь, мировых, изучение их не может полностью обеспечить многосторонние и неуклонно расширяющиеся международные связи. Это объясняется причинами как количественного, так и качественного характера. Во-первых, огромное количество языков вовлечены в международные контакты. Во-вторых, невозможно обеспечить достаточно высокий уровень владения иностранными языками при их массовом изучении. Таким образом, возрастающая роль опосредованной коммуникации представляет собой исторически обусловленную закономерность. Современный всплеск интереса к структуре ЕЯ вызван необходимостью его использования в целом ряде перспективных научных и прикладных направлений, наибольший импульс которым придало широкое внедрение систем компьютерной обработки информации во все области деятельности.

В настоящей работе основой всех методов обработки текстов на ЕЯ является семантическая модель естественного языка, разработанная профессором факультета Прикладной математики – Процессов управления д.ф.-м.н. Тузовым В.А.

Существенным, на наш взгляд, результатом проделанной работы является получение семантического словаря английского языка [1]. Он позволяет эффективно реализовать разработанные алгоритмы поиска релевантной информации в текстах на естественном языке. В работе приведено нескольких из них - поиск, диалог и рубрицирование. Семантический анализатор, пользуясь информацией заключенной в семантическом словаре, способен извлечь из текста всю информацию,

необходимую для сколь угодно точного решения любой из названных выше задач. Системы на основе этих алгоритмов помогают быстро получить необходимые знания из больших объемов информации и гарантируют высокую степень точности результата.

Цель работы. Повышение качества использования естественного языка в компьютерных системах за счет разработки методов и средств обработки текстов на естественном языке с использованием формализованного представления ЕЯ. Построение адекватной модели семантического словаря для английского языка. Разработка методов автоматического поиска, диалога и рубрицирования информации в тексте на английском языке.

Методы исследования. В диссертации для построения логического аппарата используется математический аппарат теории формальных грамматик, теории множеств и реляционной алгебры. Для реализации поставленных целей применяются теория формализации естественных языков, в частности, функциональная модель естественного языка. Областью исследования являются математическое и программное обеспечение информационных технологий, модели и методы разработки программных средств обработки данных и знаний в естественно-языковой форме, программные инструментальные средства разработки интеллектуальных систем.

Научная новизна. В ходе исследования был построен семантический словарь английского языка, не имеющий в настоящее время аналогов в мировой практике.

Практическая ценность. Построенный в ходе исследования семантический словарь английского языка может использоваться в разнообразных системах обработки естественно-язычной информации, в управление документооборотом, в работе почтовых систем, обучающих программах, поисковых системах, программах слежения за потоками

информации, автоматизации получения информации из архивов и библиотек и т.д. Семантический анализатор, используя полученный словарь, позволяет реализовать разработанные алгоритмы поиска, диалога и рубрицирования текстов на естественном языке, дающих высокую степень адекватности результата запросу.

Апробация работы. Полученный семантический словарь английского языка был успешно (89,5% адекватности выборки запросу) использован в системе автоматического рубрицирования англоязычных текстов «Гардемарин» (Санкт-Петербург, 2004 г.).

Публикации. По теме диссертации опубликовано 3 работы [1]-[3].

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения, списка литературы и приложения. Текст диссертации изложен на 96 страницах. Список литературы содержит 118 наименований.

2. Содержание работы

Введение отражает актуальность, целевую установку и задачи исследования, направленность работы на использование в компьютерных системах, конкретизирует прикладное понятие семантики текста на естественном языке.

Первая глава содержит обзор наиболее известных подходов к формализации естественного языка и опыта использования естественного языка (ЕЯ) в компьютерных системах (КС). Приводятся модели понимания ЕЯ и обобщенные структуры лингвистических процессоров, реализующих их. Даются наиболее характерные примеры практического использования языковой информации в КС. Выявляются общие черты, достоинства и недостатки обсуждаемых подходов. Завершается глава указанием наиболее перспективного подхода к формализации ЕЯ, постановкой целей и задач исследования.

Во **второй главе** содержится описание выбранного подхода к формализации ЕЯ. В первой части главы приводятся общие требования к формализованному представлению ЕЯ, излагаются основополагающие тезисы [1], лежащие в основе выбранного подхода.

Тезис 1. Язык представляет собой алгебраическую систему, $\{f_1, f_2, \ldots f_n, M\}$, где f_i – ***базисные функции*** на языке, а M – структура языка, представляющая собой набор ***базисных понятий*** $m_1, \ldots m_r$ и их иерархию (см. §2).

Тезис 2. Каждое предложение языка можно представить в виде суперпозиций базисных функций f_i, через которые выражается и каждое слово языка, за исключение базисных понятий $m_j \subset M$.

Тезис 3. Каждая часть речи играет вполне определенную роль в организации синтаксической структуры предложения.

Существительные, являясь аргументами функций, образуют структуру данных языка; прилагательные – простейшие функции на существительных; глаголы – хорошо развитые функции в основном на существительных; наречия – функции на глаголах; простые предлоги – функции на существительных; сложные предлоги и союзы – функции, аргументами которых являются суперпозиции функций.

Тезис 4. Грамматика неразрывно связана с семантикой языка и представляет собой семантический словарь.

Каждое слово описывается в виде семантической формулы, состоящей из базисных функций.

Тезис 5. Усвоение языка компьютером есть построение и пополнение семантического словаря. Машина, как и человек, может использовать некоторое слово-функцию только тогда, когда знает, как присоединять к нему аргументы.

Тезис 6. Не существует языка, который невозможно точно и строго формализовать.

Функциональная модель позволяет разбить описание языка на конечное число уровней, что серьезно упрощает процесс его формализации. Трехуровневая модель естественного языка – морфосинтаксический, семантический и прагматический – полностью решает проблему его реализации на компьютере. Разбиение на три уровня определяется не абстрактными соображениями, а принципиальным различием задач, которые решаются на каждом уровне. На первом уровне обрабатывается отдельное слово, на втором – отдельное предложение, на третьем – связный текст. На первом этапе обработки текст преобразуется в последовательность выполнимых функций. На втором этапе эта последовательность выполняется.

В процессе выполнения вычисляется значение каждого слова, и слова связываются в единую конструкцию. На третьем – полученная информация отображается в базу знаний. На сегодняшний день реализованы первые два уровня модели русского языка (РЯ). Реализация третьего уровня еще не завершена, но близка к завершению. Реализация первых двух уровней означает, что построен семантический анализатор, который переводит текст с русского языка на формальный семантический язык. Решение проблемы формализации русского языка позволяет перейти к решению прикладных задач, связанных с обработкой текстов на естественном языке.

Материал представляется на примерах для русского и английского языков, со ссылками на основополагающие тезисы, но не затрагивает вопросы конкретной компьютерной реализации.

В **третьей главе** первый раздел посвящен моделированию значения единицы языка. Для достижения этой цели необходим семантический язык. Представление значений на этом языке должно содержать их толкования, на основе которых можно адекватно описать все интуитивно ощущаемые семантические связи между различными словами, предложениями, текстами.

Для успешной формализации необходимо мощное средство (формализм), легко работающее с естественным языком, не запутывая и без

того сложную его организацию [1]. На основе приведенных тезисов, выделим основные требования к такому формализму. Он должен:

- содержать толкования единиц языка – семантический словарь;
- обладать достаточной динамичностью и гибкостью, для формализации такой «живой» субстанции, как естественный язык;
- ограничивать размножение формализаций единиц языка.
- уметь выбирать правильный смысл единицы языка, адекватный ситуации, которую она описывает;

В семантическом словаре русского языка каждому из слов языка сопоставлены одна или несколько лексем. Каждая из лексем снабжена собственным обозначением и толкованием. Толкование лексемы определяет ее допустимое множество сущностей (ее значений) в зависимости от состава, вида и значения ее аргументов. Каждое такое описание включает морфологические характеристики, семантический класс, список слов и конструкций, присоединяемых этим словом и описание семантики слова с помощью семантических функций [1].

Например,

Класс $1612 Время Интервал_времени Дата

ДАТА	$1612(!Род)
ДАТИРОВАНИЕ	N%~ДАТА$1612(S0>Caus(!Тв,InсepHab(!Род,ДАТА$1612(!Ото\!Тв))))
ДАТИРОВАТЬ	N%~ДАТА$1612(Caus(!Им,InсepHab(!Вин,ДАТА$1612(!Ото\!Тв))))
ДАТИРОВАТЬСЯ	N%~ДАТА$1612(Caus(!Тв,InсepHab(!Им,ДАТА$1612(!Ото\!Тв))))
ДАТИРОВКА	N%~ДАТА$1612(S0>Caus(!Тв,InсepHab(!Род,ДАТА$1612(!Ото\!Тв))))
ОТ	N%~ДАТА$1612(Y1>Temp(Y1:,ДАТА$1612~!Когда\!Род))
ОТО	N%~ДАТА$1612(Y1>Temp(Y1:,ДАТА$1612~!Род))
ЧИСЛО	$1612(МЕСЯЦ$1603~!Род)
ПОСЛЕДНИЙ	N%~ЧИСЛО$1612(A1>Temp(A1:ЧИСЛО$1612(ВРЕМЯ$16~!Род), КОНЕЦ$11103(МЕСЯЦ$1603(!Max))))

Таким, образом, в настоящее время полное описание слова в семантическом словаре русского языка содержит следующие элементы:

- морфологические характеристики;
- семантический класс;
- семантико-грамматический тип, порождаемый словом или конструкцией;
- описание моделей управления: слова и конструкции каких типов и классов может присоединять данное слово;
- описание семантики слова в виде формулы: семантика производных слов выражается через базовые с помощью семантических функций.

Для ряда слов некоторые элементы описания отсутствуют (например, описание семантики для базовых слов).

Разделение на семантические классы необходимо для описания моделей управления слова [1]. Для каждого слова существует ряд аргументов, которые оно может присоединять. Один из способов определить эти аргументы – указать класс, которому должен принадлежать аргумент. Проблема неоднозначности слов решается простым указанием классов.

Например, слова "острый" и "коса" неоднозначны. В синтаксическом словаре для них существуют следующие альтернативы:

ОСТРЫЙ

ОСТРЫЙ $12/01905(Z0:a> ВЕЩЬ$1213\КЛЫК$124/41,Z1: !поДат,Z2: !Ото)
ОСТРЫЙ N%~0002.БОЛЕЗНЬ$124/2(Z0:a> БОЛЕЗНЬ$124/2)
ОСТРЫЙ N%~0003.ВАЖНОСТЬ$1101631(Z0:a> ИНФОРМАЦИЯ$13154,Z1: !Для,Z2: !Тв\!вПред)
ОСТРЫЙ N%~0004.ВОСПРИЯТИЕ$124/00(Z0:a> ВОСПРИЯТИЕ$124/00)
ОСТРЫЙ N%~0005.ДЕЙСТВИЕ$15(Z0:a> ДЕЙСТВИЕ$15)
ОСТРЫЙ N%~0006.ДУХ$131(Z0:a> ДУХ$131)
ОСТРЫЙ N%~0007.ОСТРОТА$13124(Z0:a> СЛОВО$1441)
ОСТРЫЙ N%~0008.ПИЩА$124/1(Z0:a> ПИЩА$124/1\ОБЕД$15205)
ОСТРЫЙ N%~0009.ПОТРЕБНОСТЬ$1303(Z0:a> ПОТРЕБНОСТЬ$1303)
ОСТРЫЙ N%~0010.СИЛЬНЫЙ$110/13(Z0:a> ХАРАКТЕРИСТИКА$12/0)
ОСТРЫЙ N%~0011.ЧУВСТВО$1300(Z0:a> ЧУВСТВО$1300)

КОСА

КОСА $1213113(Z1: !Род\!У,Z2: !Для)
КОСА $122416(Z0:s> БЕРЕГ$122416,Z1: !Род,Z2: !Где)

КОСА $1241/121(Z1: ЧЕЛОВЕК$1241~!Род\!У)

Однако, словосочетание "острая коса" становится вполне однозначным, поскольку совместимыми оказываются только альтернативы:

ОСТРЫЙ$12/01905(Z0:a>ВЕЩЬ$1213\КЛЫК$124/41,Z1:!поДат,Z2:!Ото)//001

КОСА $1213113(Z1: !Род\!У,Z2: !Для)//001

В данном случае устранение неоднозначности возможно только за счет использования семантических классов.

Существует два подхода к описанию классов. В первом случае класс определяется набором семантических признаков. При таком подходе класс может даже не указываться явно. Этот подход является более гибким и универсальным, но и значительно более сложным. Для более-менее полного словаря языка (100 - 150 тысяч слов) задание для каждого слова точного набора семантических признаков становится слишком трудоемкой задачей [1]. В нашей работе выбран другой подход, наиболее простой и эффективный: класс задается номером, отражающим его место в иерархии с одиночным наследованием. Описание таких классов собрано в специальном классификаторе.

При построении классификации мы руководствовались правилами:

- Все элементы класса должны иметь схожие семантические свойства, которые наследуются от другого класса, определяемого как надкласс. Элементы класса должны не только наследовать все семантические свойства надкласса, но могут и иметь свои индивидуальные признаки. Класс, стоящий в корне этой древовидной структуры назвали «нечто». Один или несколько подклассов могут иметь многие классы, но слишком мелкое деление является нецелесообразным.
- Если один элемент какого-либо класса может использоваться как аргумент в некоторой семантической формуле, то и остальные

элементы этого класса могут быть аргументами в этой же формуле. То же должно выполняться и для подклассов данного класса, причем обратное не всегда верно.

Каждому из классов присваивается свой идентификационный номер, причем желательно, чтобы по номеру можно было определить надкласс данного класса. Настоящий способ иерархического описания является расширяемым.

Семантический словарь русского языка - мощный инструмент для решения многих вопросов обработки текстов. На основе информации, заключенной в нем, можно построить формальное представление текста с помощью семантического анализатора В. А. Тузова [1, 2]. Анализатор решает две проблемы: правильный выбор (как правило, единственной) альтернативы слова и связывание выбранных альтернатив в единую конструкцию.

Проанализированный таким способом текст можно использовать в качестве входного для систем распознавания текстов, информационного поиска, классификации и рубрикации, синтеза текстов, реферирования и аннотирования, диалога и даже машинного перевода. Перечисленные задачи эффективно решаются для русского языка, и в настоящее время имеются несколько экспериментальных действующих систем.

Для работы с англоязычными текстами и создания эффективных систем их обработки было принято решение построить семантический словарь для английского языка (АЯ). При построении структуры подобной уже существующему семантическому словарю русского языка мы использовали семантический словарь русского языка и электронную версию переводного англо-русского словаря (60 700 словарных статей английского языка) [2]. Слиянием этих двух ресурсов, был построен третий словарь, в котором каждому английскому слову сопоставлены все возможные переводы на русский язык (из переводного словаря), а каждому переводу - семантическая конструкция со всеми необходимыми характеристиками (из

семантического словаря русского языка): номером класса, морфологическими признаками, списком присоединяемых аргументов и т. д.

Полученный семантический словарь содержал около 1 млн. строк. Как видно из фрагмента, словарная статья объединяет английское слово или словосочетание, все его возможные русскоязычные переводы, далее набор семантических альтернатив для каждого слова перевода. В этой части статьи находятся лишние альтернативы, порожденные в виду неоднозначности русского языка и невозможности программно отобрать необходимые конструкции [2].

Например,

TAVERN

* таверна кабачок

(1) ТАВЕРНА $123402(!Род) {ж1 364}

(2) КАБАЧОК $122131(!Род,!Откуда) {м3 1080}

(3) КАБАЧОК N%~КАБАК$123402(Karese(КАБАК$123402(!Род))) {м3 1080}

* бар

(4) БАР $123402(!Род) {м1 12}

(5) БАР $14215/305() {м1 12}

В этой словарной статье присутствуют лишние альтернативы: (2) определяет растение, а (5) –меру давления. После удаления всех альтернатив, не относящихся к классу $123402 (ФО Поселения Учреждения Торговля_и_Обслуживание), получим

TAVERN

* таверна кабачок

ТАВЕРНА $123402(!Род) {ж1 364}

КАБАЧОК N%~КАБАК$123402(Karese(КАБАК$123402(!Род))) {м3 1080}

* бар

БАР $123402(!Род) {м1 12}

Такой тщательный анализ был применен ко всему словарю и в настоящее время существует окончательная его версия, содержащая 969 816 строк. Созданием этого словаря мы добились одной из наших **основных**

задач: мы отобразили подмножество английских слов (ограниченное количеством словарных статей англо-русского переводного словаря) на множество семантических классов РЯ. Вопрос о соответствии семантической классификации (классов) явлений и предметов для русского и английского языков может показаться спорным из-за различий в культурно-традиционном аспекте. Однако для необходимого в нашем исследовании уровня работы с англоязычными текстами семантические классы РЯ могут считаться адекватными и для АЯ.

Полученный семантический словарь английского языка открывает широкие перспективы для обработки англоязычных текстов, в частности поиска, диалога на уровне «вопрос–ответ» и рубрикации.

Во втором разделе главы описываются алгоритмы, разработанные для обработки англоязычных текстов, использующих в качестве входной информации формальное представления текста на естественном языке, формируемое семантическим анализатором на основе семантического словаря [3].

Алгоритм *поиска информации* в тексте: каждое предложение текста переводится в вектор-предложение, содержащее семантические классы, к которым относится каждое слово предложения. При этом в вектор войдет тот класс, к которому относится слово именно в этом предложении. Определить эти классы позволяет построенный семантический словарь для английского языка. Слова запроса переводятся в форму вектора-запроса. Далее каждое вектор-предложение и запрос проверяются на совпадение классов. Для задания точности выборки вводится рейтинг предложения – количество совпавших семантических классов. Задавая рейтинг можно получить результат требуемой точности. Такой механизм поиска позволяет отобрать не только те предложения, которые непосредственно содержат слова запроса, но и те, в которых есть близкие по смыслу (относящиеся к одному

семантическому классу), что существенно повысит процент адекватности выборки запросу.

Диалог на уровне, когда на запрос пользователя система реагирует определенным образом, можно реализовать с помощью семантического словаря. Каждое предложение текста, в котором содержатся ответы на наши будущие вопросы, переводится в вектор – цепочку семантических классов, упорядоченных, например, по возрастанию. Каждый аргумент этого вектора – семантический класс, к которому относится слово данного предложения. Получаемая структура является глубоко информативной. Набор таких векторов можно называть базой знаний. Запрос к базе знаний строится на обычном естественном языке. На следующем этапе запрос приобретает форму вектора из семантических классов, и из текста выбираются все те предложения, в которых произошли совпадения по аргументам (семантическим классам). Чем больше аргументов совпадает, тем больше данное предложение будет удовлетворять нашему запросу. Для того чтобы выбрать наиболее правильный ответ на запрос (отвечающий смыслу), необходимо выбрать те предложения, вектора которых совпали по максимальному количеству позиций с нашим запросом. Для этого мы вводим рейтинг предложений, который вычисляется при проверке каждого вектора предложения на совпадение с вектором запроса. После этого, исследуя рейтинг, легко получаем ответ на наш запрос с высокой степенью точности.

Рубрикация на основе семантического словаря способна дать очень высокую точность. Предлагаемая система позволяет осуществлять гибкую настройку глубины и направления рубрикации в соответствии с требованиями заказчика. Каждое предложение текста преобразуется в набор номеров классов (в обобщенном варианте – каждый текст). Он является средством определения степени близости двух предложений (текстов). Степень близости определяется количеством совпадающих классов. Чем больше совпадает классов, тем ближе они по смыслу. Рубрикация строится

следующим образом: для каждой пары предложений (текстов) определяем степень близости, далее получаем таблицу из чисел-степеней. Выбираем из нее максимально совпадающие и получаем нижний уровень иерархии рубрикатора. Затем определяем степень близости сгруппированных текстов и до тех пор, пока мы не получим все возможные варианты в виде древовидной структуры. Каждая группа этой структуры представляет собой предложения (тексты) относящиеся к одной тематике. Таким образом, научившись определять степень семантической близости, мы можем реализовать задачу рубрикации (по существу авторубрикации) с высокой степенью точности.

В приложении приводятся таблица соответствий некоторых лингвистических терминов, примеры подходов к формализации естественного языка, классификатор семантических классов, примеры поиска, диалога и рубрицирования, фрагмент словарных статей семантического словаря английского языка.

3. Основные результаты работы

1. Доказана эмпирическим путем возможность адекватного отображения подмножества английских слов на множество семантических классов русского языка.
2. Создан семантический словарь английского языка, где каждому английскому слову сопоставлены все возможные переводы на русский язык, а каждому переводу - семантическая конструкция со следующими характеристиками: номером семантического класса, морфологическими признаками, списком присоединяемых аргументов и т. д. На основе этих атрибутов семантический анализатор способен автоматически построить формальное представление текста.

3. Разработаны методы обработки англоязычных текстов на основе созданного семантического словаря АЯ, в частности, поиска, диалога и рубрицирования.

4. Публикации автора по теме диссертации

[1] Кутарба А.Ю. Семантический словарь для естественного языка. // «Актуальные проблемы науки в России». Материалы Всероссийской научно-практической конференции. Вып.3. Т. 2. — Кузнецк, 2005. — С.228-232.

[2] Кутарба А.Ю. Особенности построения семантического словаря английского языка. Деп. В ВИНИТИ № 1734 от 26.12.2005, 12 с.

[3] Кутарба А.Ю. Обработка англоязычных текстов на основе семантического словаря. // Вестник С.-Петерб. ун-та. Сер.1. 2005. Вып.3-4. С.46-53.