Аннотация

Диссертационная работа посвящена изучению приложений *функциональной теории естественного языка* и автоматического *семантического анализатора* проф. В.А. Тузова. Являясь мощным инструментом исследования структуры предложений русского языка и выявления смысла отдельных слов, семантический анализатор почти не применяется на практике. В диссертации продемонстрированы алгоритмы решения некоторых задач из области обработки текстов на естественном языке, опирающиеся на извлекаемую семантическим анализатором информацию.

Общая характеристика работы

Актуальность. Обработка текстов на естественном языке — важная задача, привлекающая внимание специалистов на протяжении десятилетий. Анализ документов требуется, в частности, для решения проблем информационного поиска, машинного перевода, функционирования вопросно-ответных систем и модулей проверки правописания. Семантический анализатор может существенно улучшить качество систем, связанных с обработкой текстов, но способы его использования пока ещё слабо изучены.

Цель работы — изучить возможность применения семантического анализатора в реальных проектах. Для этого требуется разработать ряд алгоритмов, опирающихся не на классические модели представления знаний документов (основанные на статистике, либо на поверхностном анализе), а на деревья разбора предложений и синтактико-семантические описания слов, генерируемые семантическим анализатором. Построенные

алгоритмы должны быть запрограммированы на уровне, по крайней мере, экспериментальных систем.

Направления исследований:

1. Сравнение функциональной модели языка, предложенной В. Тузовым, с более ранними теориями.
2. Изучение возможности использования функциональной модели языка и семантического анализатора в задачах построения вопросно-ответных систем, информационного поиска и рубрикации, проверки правописания и подбора синонимов, поиска частичных совпадений и выявления плагиата, а также в машинном переводе.
3. Анализ технических деталей, связанных с применением семантического анализатора в реальных проектах.

Методы исследования, достоверность и обоснованность результатов. Предлагаемая работа ориентирована на достижение практических результатов. Почти все описанные алгоритмы доведены до реализации, их работоспособность подтверждается экспериментами. Теоретические построения (классификация вопросительных предложений, схема системы автоматизированного перевода) опираются на известные, описанные в научной литературе результаты. Фундаментом исследований служит функциональная модель языка В. Тузова, цитируемая во многих научных работах.

На защиту выносятся:

1. Предложенные способы использования семантического анализатора В. Тузова в различных проектах, связанных с обработкой текстов на естественном языке.

2. Конкретные алгоритмы, лежащие в основе построенных экспериментальных систем.
3. Теоретические модели, такие как классификация вопросительных предложений в русском языке и общая схема системы автоматизированного перевода.

Научная новизна результатов исследования:

1. Впервые была широко изучена возможность использования функциональной модели языка и семантического анализатора для решения практических задач, связанных с обработкой текстов на естественном языке.
2. Автором разработаны конкретные алгоритмы решения ряда задач, относящихся к теме исследования. Построены экспериментальные системы, иллюстрирующие выполнение алгоритмов семантического анализа.
3. Автором создана модель системы автоматизированного перевода, использующей функциональную теорию языка.

Практическая полезность работы. Предлагаемые в диссертации методы могут использоваться при создании высококачественных систем обработки текстов на естественном языке. Семантический анализатор не имеет аналогов по качеству и полноте генерируемых выходных данных. Отдельные элементы выходной распечатки могут быть также применены в существующих программных продуктах для реализации дополнительной функциональности.

Реализация результатов работы. Почти все описываемые в работе алгоритмы воплощены в экспериментальных системах, предназначенных для решения задач, относящихся к обработке документов на естественном

языке. Функционирование данных систем подробно описывается в научных работах автора.

Апробация работы. Отдельные результаты по теме диссертации докладывались:

1. на XXXVII конференции «Процессы управления и устойчивость» (С.-Петербург, 10-13 апреля 2006г.);
2. в Летней школе IMPDET (Мекриярви, Финляндия, 4-9 июня 2006г.);
3. на специальном семинаре кафедры технологии программирования факультета ПМ-ПУ СПбГУ;
4. в рамках цикла лекций представителей факультета ПМ-ПУ СПбГУ в университете г. Аизу (Япония, 19-24 февраля 2006г.);
5. на семинаре PhD студентов кафедры компьютерных наук факультета естественных наук университета г. Йоэнсуу (Финляндия).

Публикации. Основные результаты диссертации отражены в 3 научных работах; результаты ещё 4 работ существенно используются при решении отдельных изучаемых задач.

Структура и объём работы. Диссертационная работа состоит из введения, семи глав, заключения и библиографического списка, включающего 63 наименования. Работа изложена на 116 листах машинописного текста, содержит 15 рисунков и 24 таблицы.

Краткое содержание работы

Первая глава знакомит читателя с формальными моделями естественного языка. Попытки строго научного описания языков предпринимаются, по крайней мере, с пятидесятых годов XX века (если не считать единичных работ XIX столетия и даже более раннего времени).

Лишь немногие из них, однако, оказали существенное влияние на современное состояние NLP. Мы рассмотрим три возможных подхода: грамматики Хомского как наиболее влиятельную модель, оказавшую большое воздействие на теорию компиляции, модель «смысл ó текст» И. Мельчука, охватывающую самые разные пласты языкознания, и функциональную теорию языка В. Тузова, на основе которой был разработан семантический анализатор. Теории, посвящённые частным аспектам языка (морфологии, синтаксису) в работе не рассматриваются.

Вторая глава иллюстрирует, как семантический анализатор может быть применён в задаче разработки вопросно-ответных систем, предназначенных для организации полноценного интерфейса на естественном языке между человеком и компьютером. Во второй главе также рассматривается классификация вопросительных предложений, имеющих смысл в контексте диалога с компьютером.

Третья глава посвящена задачам информационного поиска и рубрикации документов. Современные системы поиска и рубрикации обычно основываются на статистическом анализе текстов и анализе различных эвристических показателей (таких как популярность документа и количества ссылок на него, если речь идёт о странице в интернете). Это делает используемые алгоритмы независимыми от языка документов, но не позволяет использовать информацию, напрямую заложенную в слова. Семантический анализатор способен сделать поиск более интеллектуальным, что доказывается на примерах применения *словаря классов* и *деревьев разбора предложений*.

В четвёртой главе описывается механизм использования семантического анализатора в задачах проверки правописания и подбора синонимов слов. Семантический анализатор основан примерно на тех же

принципах, что и компилятор языка программирования, поэтому (в частности) проверка правильности структуры входных предложений является его прямой задачей. Кроме того, в состав анализатора входит *семантический словарь*, которым можно воспользоваться как словарём синонимов.

В пятой главе рассматривается задача поиска частично совпадающих документов и выявления плагиата. Алгоритмы, разработанные для её решения, оказываются особенно эффективными при обработке информации, имеющей некоторую структуру. Неструктурированные данные приходится сравнивать достаточно простыми средствами, в то время как для файлов, поддающихся структурному анализу, можно создать более качественную специализированную процедуру. Семантический анализатор способен структурировать тексты на естественном языке, расширяя возможности для разработки эффективных алгоритмов их сравнения.

В шестой главе изучается возможный подход к решению задачи машинного перевода с помощью семантического анализатора. Машинный перевод изобилует неожиданными трудностями, поэтому говорить о возможности полноценного его осуществления с помощью применения какой-либо технологии не приходится. Однако принципы, на которых основан семантический анализатор, позволяют естественным образом решать задачи, оказывающиеся весьма сложными для других методов построения автоматизированных систем перевода.

Седьмая глава фокусирует внимание на некоторых технических аспектах, связанных с использованием семантического анализатора. Анализатор представляет собой сложную систему, предназначенную для решения нетривиальных задач, и способ его общения с внешним миром

сам по себе заслуживает внимания. Также здесь обсуждаются перспективы развития семантического анализатора как программного продукта.

Заключение

Целью данной работы была попытка показать, что семантический анализатор может быть применён при решении самых различных задач, связанных с обработкой текстов на естественном языке. На нынешний момент нам представляется, что именно широта охвата предметной области могла бы привлечь внимание к алгоритмам семантического анализа и помочь понять, где анализатор может быть особенно эффективен.

В рамках исследований изучались такие направления, как создание вопросно-ответных систем, информационный поиск и рубрикация, инструменты проверки правописания и подбора синонимов, поиск частичных совпадений и выявление плагиата, а также машинный перевод. Были разработаны:

- экспериментальная вопросно-ответная система первого уровня понимания;
- классификация вопросительных предложений, пригодная для последующего использования в диалоговых приложениях;
- система информационного поиска, опирающаяся на семантические формулы слов документов коллекции;
- модуль поиска связанных слов;
- контекстно-ориентированный электронный тезаурус;
- система поиска плагиата в текстах на русском языке, использующая систему классов как основу модуля токенизации;

§ рабочая модель системы машинного перевода.

Список основных публикаций

[1] Мозговой М.В. Простая вопросно-ответная система на основе семантического анализатора русского языка // Вестник СПб университета. — 2006. — сер. 10. — вып. 1. — С. 116-122.

[2] Мозговой М.В. Семантический анализатор и задача информационного поиска // Вестник СПб университета. — 2005. — сер. 10. — вып. 3. — С. 54-59.

[3] Мозговой М.В. Контекстно-ориентированный тезаурус русского языка // Процессы управления и устойчивость: Труды 37-й международной научной конференции аспирантов и студентов / Под ред. А.В. Платонова, Н.В. Смирнова — СПб.: Изд-во СПбГУ. — 2006. — С. 379-383.

Список сопутствующих публикаций

[4] Mozgovoy M. Desktop Tools for Offline Plagiarism Detection in Computer Programs // Informatics in Education. — 2006. — Vol. 5(1). — P. 97-112.

[5] Fredriksson K., Mozgovoy M. Sublinear Parameterized Single and Multiple String Matching. Technical Report A-2006-2, Department of Computer Science, University of Joensuu, March, 2006.

[6] Mozgovoy M., Fredriksson K., White D., Joy M., and Sutinen E. Fast Plagiarism Detection System // Lecture Notes in Computer Science. — 2005. — Vol. 3772. — P. 267-270.

[7] Мозговой М.В. Классика программирования: алгоритмы, языки, автоматы, компиляторы. Практический подход. — СПб.: Наука и Техника, 2006. — 320 с.