В.А. Фомичев

ФОРМАЛИЗАЦИЯ ПРОЕКТИРОВАНИЯ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

МАКС ПРЕСС

МОСКВА 2005

В монографии описывается апробированная на практике новая система взаимосвязанных формальных моделей и алгоритмов, предназначенных для проектирования лингвистических процессоров (компьютерных систем, осуществляющих смысловую обработку письменных текстов или устной речи на естественном языке) в произвольных предметных областях. Значительное внимание уделяется изложению оригинального теоретического подхода к математическому описанию смысловой структуры не только предложений, но и сложных связных текстов (или дискурсов), относящихся к деловой прозе: текстов по медицине, экономике, юриспруденции и т.д. Анализируются возможности использования этого подхода в теории многоагентных систем, для разработки логико-информационных основ электронной коммерции и для устранения языкового барьера между пользователями сети Интернет из разных стран.

Значительная часть материалов монографии была опубликована в научных журналах “Информационные технологии”, “Качество и ИПИ (CALS)- технологии”, “Качество. Инновации. Образование”, “Informatica” (Словения), “Cybernetica” (Бельгия) и трудах международных научных конференций и симпозиумов, проходивших в России, Австрии, Великобритании, Германии, Дании, Нидерландах, Словении, Франции.

Книга не имеет аналогов в мировой научной литературе и будет полезна как опытным специалистам в области прикладных интеллектуальных систем или математической лингвистике, так и студентам и начинающим ученым.

The monograph describes a new system of interrelated formal models and algorithms tested in practice and destined for designing linguistic processors, or natural language processing systems (computer systems fulfilling the conceptual processing of written texts or oral speech in natural language) in arbitrary application domains. A considerable attention is drawn to setting forth an original theoretical approach to representing in a mathematical way the structured meanings (or conceptual structure, semantic structure) of not only separate sentences but also of complicated narrative texts (or discourses) pertaining to medicine, economy, law, and other fields of professional activity. The possibilities of using this approach in the multi-agent theory, for the elaboration of logical-informational foundations of electronic commerce (e-commerce), and for the elimination of the language barrier between the Internet users from various countries are analysed.

A considerable part of the stated materials was published in the scientific journals in Russian “Informational Technologies”, “Quality and IPI (CALS)-Technologies”, “Qulaity. Innovations. Education”, in the international scientific journals “Informatica” (Slovenia), “Cybernetica” (Belgium), and in the proceedings of the international scientific conferences and symposia which were held in Russia, Austria, Denmark, France, Germany, Slovenia, The Netherlands, and United Kingdom.

The monograph has no analogues in the world scientific literature and will be of use both to experienced specialists in the field of applied intelligent systems or mathematical linguistics and to the students and young scientists.

ОГЛАВЛЕНИЕ

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Всегда практика должна быть воздвигнута на
хорошей теории, ворота которой - перспектива
Леонардо да Винчи

В преподавании такой быстро развивающейся области,
какой является наука о вычислительных процессах,
правильный педагогический принцип состоит в том,
чтобы больше внимания уделять идеям, а не техни-
ческим подробностям реализации
А. Ахо, Дж. Ульман

За последние два десятилетия научно-техническое направление "искусственный интеллект" получило значительное развитие и нашло целый ряд успешных применений. Основная часть информации хранится и передается людьми с помощью естественного языка (ЕЯ), т.е. совокупности русского, английского, японского и других языков. Один из главных подклассов компьютерных систем с элементами искусственного интеллекта (СИИ) составляют программы, понимающие ЕЯ или синтезирующие выражения ЕЯ по некоторым внутренним представлениям. Такие программы называются системами обработки естественного языка (в англоязычной научной литературе: natural language processing systems), или лингвистическими процессорами (ЛП). Технологии, предусматривающие использование ЛП для обработки информации, составляют основной подкласс лингвистических информационных технологий (ЛИТ).

Другие виды современных ЛИТ связаны с разработкой и применением языков общения компьютерных интеллектуальных агентов (КИА) в многоагентных системах, языков построения протоколов переговоров, проводимых КИА в области электронной коммерции, и языков формирования контрактов, заключаемых КИА в ходе таких переговоров, а также семантически-структурированных языков нового поколения для представления информации во Всемирной Паутине (the World Wide Web, или WWW).

Несколько неформальных понятий, являющихся базовыми для теории смысловой обработки компьютером естественного языка, многократно используются в этой книге: семантика естественного языка, связный текст (или дискурс), структурированное значение выражения на ЕЯ, семантическое представление ЕЯ-выражения и алгоритм семантико-синтаксического анализа.

Под семантикой ЕЯ будем понимать совокупность закономерностей передачи информации средствами ЕЯ. Связным текстом (или дискурсом) называется последовательность взаимосвязанных по смыслу выражений на ЕЯ.

Если Т – некоторое выражение на ЕЯ (словосочетание, предложение, дискурс), то структурированным значением выражения Т является информационая структура, строящаяся мозгом человека, владеющего данным подязыком ЕЯ (русским, английским или другим), независимо от контекста, в котором услышано или прочитано выражение Т, т.е. строящаяся на основе только знаний о значениях элементарных лексических единиц и правил их комбинирования в данном языке.

Под семантическим представлением (СП) ЕЯ-выражения Т понимается формальная структура, являющаяся либо образом структурированного значения этого выражения, либо отражением смысла (или содержания) данного выражения в определенном контексте - в конкретной ситуации диалога, в контексте знаний о мире или в контексте предшествующей части дискурса.

Таким образом, СП ЕЯ-выражения Т является формальной структурой, первичными элементами которой являются, в частностия, обозначения понятий, конкретных объектов, множеств объектов, событий, имена функций и отношений, логические связки, обозначения чисел и цветов, а также обозначения смысловых отношений между значениями фрагментов текста или между объектами рассматриваемой предметной области.

СП текстов могут являться, например, строками и размеченными ориентированными графами (семантическими сетями).

Алгоритм семантико-синтаксического анализа строит по тексту на ЕЯ его СП, используя для этого знания о морфологии и синтаксисе подъязыка ЕЯ (русского, английского и др.), информацию о взаимосвязях лексических единиц с единицами семантического уровня и знания о мире. Семантическое представление текста, построенное таким алгоритмом, интерпретируется прикладной интеллектуальной

системой в зависимости от ее назначения, например, как задание на поиск ответа на вопрос, команда на выполнение физического действия автономным интеллектуальным роботом, фрагмент знаний о мире, предназначенный для пополнения базы знаний и т.д.

Научные результаты, изложенные в данной монографии, были получены автором в ходе цикла исследований, начатого более двадцати лет назад. Выбор направления исследований был реакцией на почти полное отсутствие в то время эффективных математических средств и методов проектирования ЛП.

Результаты данной монографии дают не только продвижение вперед, но и *качественный скачок* в области разработки формальных средств и методов проектирования алгоритмов семантико-синтаксического анализа ЕЯ-текстов. Этот качественный скачок обусловлен следующими основными факторами:

1. Разработчики ЛП получили систему правил (причем компактную, состоящую всего из 10 основных правил), позволяющих, по гипотезе автора, строить семантические представления произвольных текстов деловой прозы, т.е. текстов по экономике, технике, медицине, юриспруденции и т.д. Это означает, что эффективные процедуры построения СП ЕЯ-текстов и процедуры обработки СП ЕЯ-текстов (в контексте содержания предшествующей части текста или диалога, в рамках знаний о предметной области и т.д.) можно будет использовать в разных предметных областях и развивать возможности этих процедур при возникновении новых задач.
2. Построена формальная модель лингвистической базы данных, содержащей такие сведения о лексических единицах и их взаимосвязях с информационными единицами, которые достаточны для семантико-синтаксического анализа интересных для приложений подъязыков русского языка.
3. Разработан практически полезный сложный структурированный алгоритм семантико-синтаксического анализа, который описывается не средствами какой-либо системы программирования, а полностью с помощью предложенной системы формальных понятий, что делает этот алгоритм независимым от программной реализации и предметной области.

СОДЕРЖАНИЕ КНИГИ

В главе 1 дается краткий обзор областей применения лингвистических процессоров, а также анализируются потребности расширения запаса эффективных формальных средств и методов для проектирования ЛП и разработки ЛИТ в области многоагентных систем и электронной коммерции.

В главе 2 описывается математическая модель, перечисляющая первичные единицы концептуального уровня, используемые ЛП, а также описывающая информацию, связанную с такими единицами и необходимую для соединения этих единиц в составные единицы, отображающие структурированные значения (СЗ) сколь угодно сложных ЕЯ-текстов.

В главе 3 (в развитие результатов главы 2) построена математическая модель для описания СЗ предложений и сложных связных текстов (дискурсов) на естественном языке (в частности, на русском, английском, немецком, французском языках). Модель представляет собою определение нового класса формальных языков, названных стандартными концептуальными языками (стандартными К-языками, СК-языками), и может рассматриваться как формальная грамматика нового вида. Сущность этой модели в том, что она задает 10 операций на концептуальных структурах, с помощью которых за конечное число шагов можно построить семантическое представление предложения или дискурса из чрезвычайно широкого подъязыка деловой прозы.

Проведено математическое исследование формальных объектов, задаваемых этой моделью – выражений СК-языков. В частности, доказана однозначность структурного анализа таких выражений.

Глава 4 посвящена исследованию выразительных возможностей класса СК-языков. Показано, что выражения СК-языков удобно использовать для: (а) построения СП предложений (выражающих высказывания, вопросы, команды) и сложных дискурсов на русском языке, (б) построения составных целей, (в) представления знаний о мире, в том числе для построения формальных определений понятий и объектно-ориентированных модулей знаний..

Проведено сравнение выразительных возможностей СК-языков с выразительными возможностями других, наиболее часто используемых подходов к

формальному представлению значений (смысловой структуры) ЕЯ-текстов: теории представления дискурсов, теории концептуальных графов, эпизодической логики, теории расширенных семантических сетей, теории неоднородных семантических сетей и компьютерной семантики русского языка. Показано, что выразительные возможности СК-языков значительно превосходят возможности перечисленных подходов и, в то же время, аппарат СК-языков позволяет моделировать механизмы представления информации, характерные для каждого из указанных подходов.

В главе 5 исследуются возможности использования аппарата СК-языков для решения ряда актуальных проблем информатики: разработки языков представления содержания посланий компьютерных интеллектуальных агентов, в частности, языков, предназначенных для формирования контрактов и протоколов переговоров в области электронной коммерции, создания семантического сетевого языка нового поколения, построения онтологий предметных областей, разработки новых языков представления знаний для решения информационно-сложных задач, проектирования интеллектуальных информационно-поисковых и вопросо-ответных Интернет-систем нового поколения.

В главе 6 вводится формальное понятие лингвистического базиса, которое интерпретируется как описание структуры лингвистической базы данных (ЛБД), используемой алгоритмом семантико-синтаксического анализа ЕЯ-текстов. ЛБД, структура которых отображается построенной моделью, позволяют устанавливать возможные смысловые отношения, в частности, в сочетаниях «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение или Наречие + Глагол», «Предлог + Вопросительно-относительное местоимение + Глагол».

В главе 7 излагается новый метод преобразования ЕЯ-текстов в их семантические представления. Метод предусматривает использование предложенного автором матричного семантико-синтаксического представления (МССП) входного текста как промежуточного представления при переходе от ЕЯ-текста к СП текста, являющемуся выражением некоторого СК-языка (т.е. К-представлением текста).

При этом не используется традиционное синтаксическое представление текста. Тексты могут быть, в частности, вопросами, сообщениями (описаниями фактов, ситуаций) или командами.

В главах 8 и 9 разработан сложный структурированный алгоритм семантико-синтак-сического анализа текстов из представляющих практический интерес подъязыков естественного (русского) языка (алгоритм SemSyn). Этот алгоритм, базирующийся на построенной в главе 6 формальной модели ЛБД и на введенном в главе 7 понятии МССП текста, устанавливает смысловые отношения между элементарными значащими единицами входного текста, отражая эти отношения посредством МССП, а затем строит СП текста, являющееся выражением некоторого СК-языка (К-представлением). Входные ЕЯ-тексты могут выражать высказывания (сообщения), команды, специальные вопросы (т.е. вопросы с вопросительными словами), общие вопросы (т.е. вопросы с ответом «Да»/ «Нет»)и могут, в частности, включать причастные обороты и придаточные определительные предложения. Алгоритм SemSyn позволяет устанавливать возможные смысловые отношения, в частности, в сочетаниях перечисленных выше видов.

В заключении к данной монографии делается вывод о том, что совокупность научных результатов, изложенных в главах 1 - 4, 6 - 9, и часть научных результатов главы 5 образуют новую теорию проектирования семантико-синтаксических анализаторов естественно-языковых текстов с использованием формальных средств представления входных, промежуточных и выходных данных; эта теория может быть названа теорией К-представлений.

Приложение содержит доказательства двух лемм и базирующегося на них доказательства одного из утверждений из главы 3. Нумерация утверждений сквозная внутри каждой главы (Утверждение 3.1, Утверждение 3.2 и т.д.).

В основе большей части содержания данной монографии лежат циклы лекций, читавшиеся автором с 1996 г. студентам Российского государственного технологического университета им. К.Э Циолковского – “МАТИ” по дисциплинам “Теоретические основы лингвистических информационных технологий”, “Математическая лингвистика”, “Проектирование лингвистических процессоров” и студентам Московского государственного института электроники и математики

(технического университета) по дисциплинам “Лингвистические информационные технологии”, “Проектирование лингвистических процессоров” и “Глобальные информационные сети и дистанционное обучение”.

БЛАГОДАРНОСТИ