Проблемы бионики
1998. Вып. 49. С. 54—58

ХТУРЭ

УДК 519.711.3

О.В. АНДРИАНИ, Е.М. РОНИН, В.А. ЧИКИНА

КОМПЬЮТЕРНЫЙ ГРАММАТИЧЕСКИЙ СЛОВАРЬ

За последнее десятилетие мощность и быстродействие выпускаемых компьютеров возросли более чем на порядок, резко увеличился и объем выпуска коммерческого программного обеспечения. Однако общие концепции разработки пользовательского интерфейса программ изменились мало. Нынешние программы общаются с пользователем на том же концептуальном уровне, что и десять лет назад. Изменилось в основном оформление, улучшились визуальные характеристики (утвердился графический пользовательский интерфейс, развились Интернет-технологии).

Наиболее удобной системой диалога человека с компьютером было бы, очевидно, применение естественного языка. Разработки как речевых, так и письменных диалоговых систем ведутся уже давно, и результаты их свидетельствуют о возможности построения качественно новых систем анализа естественноязычных текстов. Но зачастую подобные системы ориентируют лишь на специальные области знания; их авторы ограничивают словарь и отказываются от возможности его расширения в пользу быстродействия и скорости разработки. В данной работе представлен прототип универсальной системы для работы с грамматической информацией на языках с развитым словоизменением.

Для проведения аналитической работы с текстами больших объемов на естественном языке был разработан программный комплекс "Компьютерный грамматический словарь", ориентированный на хранение, анализ и обработку грамматической информации.

Рассмотрим функции его модулей.

Модуль Storehouse обеспечивает управление базой данных (БД) и позволяет производить основные изменения в ее файлах — такие, как вставка и добавление записей, обновление, сортировка и другие служебные операции. Этот модуль обеспечивает высокую производительность и быстродействие программного комплекса в работе с БД.

Модуль программного интерфейса Linx формирует запросы, которые затем передаются для выполнения в модуль Storehouse. Запросы представлены в виде промежуточных таблиц БД, которые содержат исходную информацию при пополнении БД и промежуточный результат поиска. Такая схема передачи данных между приложением и программным комплексом

позволяет упростить алгоритм обработки запросов и унифицировать потоки данных.

Модуль Anword обеспечивает анализ поступающей грамматической информации и формирует запросы на ввод-вывод словоформ и парадигм словоизменения. Этот модуль может быть изменен и дополнен в различных пользовательских приложениях, что позволяет сделать обработку информации гибкой и предоставляет возможность создания многоязычных и специализированных приложений.

Модуль интерфейса Userface является необязательным элементом программного комплекса и может корректироваться при подключении различных приложений. Интерфейс пользователя состоит из визуальной системы пополнения БД, включающей в себя формы ввода для всех частей речи русского языка, и независимого пользовательского интерфейса приложения.

БД, с которой работает программный комплекс, состоит из четырех таблиц, содержащих списки основ слов и окончаний, а также матрицу связей между ними и соответствующими грамматическими признаками.

Таблица 1

Номер поля	Название поля	Размерность поля
1	Номер основы	4-байтовое число
2	Основа слова	25 символов
3	Указатель парадигмы	2-байтовое число
4	Часть речи	1-байтовое число
5	Тип словоизменения	1-байтовое число
6	Идентификационный код	4-байтовое число

Табл. 1 хранит основы слов с указанием на соответствующую парадигму, часть речи и тип словоизмения, а также идентификационный код, объединяющий слова с изменяющимися основами.

Таблица 2

Номер поля	Название поля	Размерность поля
1	Номер парадигмы	4-байтовое число
2 . . . n	Набор грамматических признаков	Необходимое количество полей размером 2 байта
n+1	Ссылка на окончание	2-байтовое число

Табл. 2 представляет собой матрицу, которая связывает конкретную парадигму с соответствующими грамматическими признаками и содержит ссылки на окончания.

Таблица 3

Номер поля	Название поля	Размерность поля
1	Номер окончания	2-байтовое число
2	Окончание	6 символов

Табл. 3 хранит окончания вместе с порядковыми номерами. Окончания расположены в таблице в алфавитном порядке и не содержат повторений.

Таблица 4

Номер поля	Название поля	Размерность поля
1	Номер окончания	2-байтовое число
2	Номер парадигмы	4-байтовое число

Кроме того, в БД включена табл. 4, которая является матрицей обратной связи. Табл. 4 содержит номера окончаний и номера соответствующих парадигм словоизменения, что позволяет осуществлять обратный переход от найденного окончания к нужной основе.

Структура БД дает возможность охватывать до 100 тысяч словоформ русского языка и является достаточно гибкой для хранения как регулярных форм, так и исключений.

В качестве тестирующего приложения в демонстрационную версию программного комплекса включена подсистема морфологического анализа русских словоформ. Для проведения подобного анализа используется алгоритм деления слова по морфологическому шву "основа — окончание". В соответствии с алгоритмом анализируемое слово делится последовательно справа налево. Истинным считается тот вариант, при котором полученная основа содержится в табл. 1 БД, отделенное окончание принадлежит списку окончаний (табл. 3), а матрица связей (табл. 2) однозначно связывает их друг с другом. После того как процесс деления завершен, найденные варианты с помощью матрицы связей сопоставляются с конкретной грамматической информацией из БД.

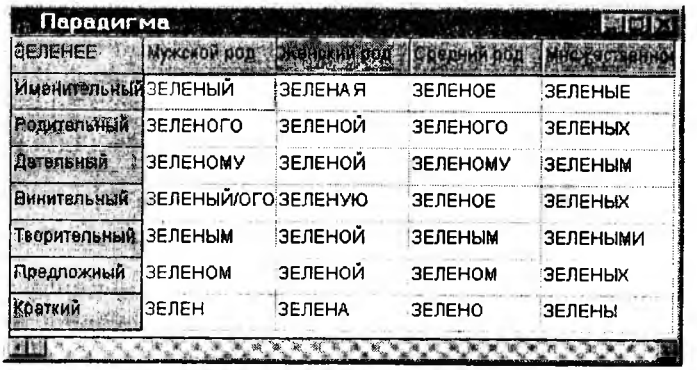
Тестирующее приложение работает в диалоговом режиме и может по желанию пользователя переключаться в альтернативные режимы представления грамматической информации — списочный и табличный.

В списочном режиме предоставляется расширенная информация, демонстрируется полный набор грамматических признаков. На рис. 1 показан

пример анализа слова «зеленому», введенного в тестирующее приложение. Элементы управления — кнопки «Назад» и «Дальше» позволяют просматривать омонимические варианты слов либо все варианты словоформ данного слова. Кнопки «Парадигма» и «Парадигмы» предназначены для перехода в табличный режим.

В табличном режиме реализуется стандартный способ отображения парадигм словоизменения. На рис. 2 представлены все формы прилагательного «зеленый».

Рис. 1



ЗЕЛЕНЕЕ	Мужской род	Женский род	Средний род	Множественное
Именительный	ЗЕЛЕНЫЙ	ЗЕЛЕНАЯ	ЗЕЛЕНОЕ	ЗЕЛЕНЫЕ
Родительный	ЗЕЛЕНОГО	ЗЕЛЕНОЙ	ЗЕЛЕНОГО	ЗЕЛЕНЫХ
Дательный	ЗЕЛЕНОМУ	ЗЕЛЕНОЙ	ЗЕЛЕНОМУ	ЗЕЛЕНЫМ
Винительный	ЗЕЛЕНЫЙ/ОГО	ЗЕЛЕНУЮ	ЗЕЛЕНОЕ	ЗЕЛЕНЫХ
Творительный	ЗЕЛЕНЫМ	ЗЕЛЕНОЙ	ЗЕЛЕНЫМ	ЗЕЛЕНЫМИ
Предложный	ЗЕЛЕНОМ	ЗЕЛЕНОЙ	ЗЕЛЕНОМ	ЗЕЛЕНЫХ
Краткий	ЗЕЛЕН	ЗЕЛЕНА	ЗЕЛЕНО	ЗЕЛЕНЫ

Рис. 2

В рамках разработки данного приложения создана усовершенствованная система ввода информации в БД, позволяющая вводить ее в автоматизированном и ручном режиме.

Работа с программным комплексом "Компьютерный грамматический словарь" не требует специальных навыков от пользователя, знакомого с интерфейсом Microsoft Windows (tm). Благодаря гибкой системе диалога с пользователем, программный комплекс позволяет легко получать и вводить информацию.

Для функционирования программного комплекса необходим персональный компьютер типа IBM PC AT с процессором 80386 или выше, обладающий не менее чем 4 Мбайт оперативной памяти, видеоадаптером и монитором VGA. Компьютер должен иметь как минимум 1 Мбайт свободного дискового пространства для установки программного комплекса.

Компьютерный грамматический словарь предназначен для работы в среде Microsoft Windows (tm) 3.1 и выше (тестирован в среде MS Windows 3.1, Win'95, Windows NT 4.0). Для функционирования программного комплекса необходимо установить драйверы Borland Database Engine.

Программный комплекс написан для среды компилятора Borland Delphi Client/Server 1.0 и отлажен в ней. Эта система выбрана в связи с тем, что она обладает широким набором средств работы с БД. Кроме того, данная версия создает 16-битные Windows-приложения, работающие в среде MS Windows версии 3.1 и выше.

В результате создания Компьютерного грамматического словаря открываются возможности проведения обширной статистической работы по определению регулярных синтаксических и смысловых конструкций естественноязычных текстов. Такого рода исследования необходимы и для теоретических разработок, посвященных сложным вопросам языкознания, и для решения прикладных задач. Приложения, созданные на базе программного комплекса, можно использовать для идентификации авторства, классификации по предметным областям, а также для построения гибких систем автоматического перевода. Появляется возможность создания специализированных словарей и приложений, предназначенных для автоматизированного аннотирования и реферирования текстов.

Список литературы: 1. *Зализняк А.А.* Грамматический словарь русского языка: Словоизменение. М.: Рус. яз., 1980. 880 с. 2. *Бондаренко М.Ф., Осыка А.Ф.* Автоматическая обработка информации на естественном языке. К.: Учеб.-метод. каб. высш. образования, 1991. 144 с. 3. *Мартин Дж.* Организация баз данных в вычислительных системах: Пер. с англ. М.: Мир, 1980. 662 с.

Поступила в редколлегию 23.01.98