А.Ф.Гельбух

ПРОСТАЯ ОБОЛОЧКА СИСТЕМЫ ТОЧНОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА И СИНТЕЗА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Настоящая работа посвящена описанию формальной модели морфологического строения естественного флективного языка (флективным языком называется язык, в грамматике которого морфологические характеристики словоформ выражаются вариацией их окончания. Русский и многие другие языки являются флективными.), а также обсуждению основанного на сформулированной модели экономичного по памяти и быстродействию алгоритма точного морфологического анализа и синтеза словоформ естественного языка. Реализованная автором пустая оболочка такой системы может быть использована для создания действующих систем для различных языков и вариантов систем для русского языка. Имеется действующая реализация для русского языка.

В настоящее время имеется опыт создания систем морфологического анализа и/или синтеза, имеющих следующие основные сферы применения: (а) в составе лингвистического процессора системы искусственного интеллекта; (б) в составе систем автоматизированной обработки текста в интеллектуальных СУБД; (в) в составе системы автоматического перевода; а также (г) в качестве средства автоматизированной диалоговой проверки грамматической правильности текста.

Анализ спектра существующих разработок и подходов показывает, что каждая такая система является замкнутым и трудоемким программным продуктом. Разнообразие таких программ объясняется разнообразием требований к ним и условий их применения и разнообразием подходов к их созданию. При этом оказывается, что, как правило, такие требования и подходы обычно можно разделить на чисто технические и чисто лингвистические. Однако, насколько известно автору, до сих пор не было предпринято серьезных попыток разделения предметной и программной информации в этой области (хотя, конечно, в каждой отдельной системе эта информация частично разделена).

Таким образом, существует острая потребность в создании взаимно устраивающей лингвиста и программиста модели предметной области, в рамках которой могли бы раздельно создаваться, отлаживаться, сопровождаться и изменяться как наборы лингвистических знаний, так и программные оболочки, подобные общеизвестным экспертным оболочкам. В докладе представлен опыт создания такой модели.

При разработке модели и основанной на ней программной оболочки особое внимание автором было уделено проблеме понимания ошибочно написанных словоформ и выдаче разумных гипотез об их исправлении. Система эффективно распознает и исправляет один класс характерных ошибок: ошибки, происходящие от выбора для слова окончания или формы, правильной для других слов языка, но не для ланного слова. Такие ошибки особенно часто делают люди, знающие основы грамматики языка, но не владеющие языком свободно. По-видимому, ценной окажется способность системы подсказывать правильное написание исключений и трудных слов, отправляясь от предложенной невозможной формы слова.

Развитый в настоящей работе аппарат в перспективе будет применен к следующим задачам: описание аномалий начертания некоторых букв, как, например, начертание русской буквы "йо" как "е"; описание расстановки ударений, что является очень важной задачей. Автор планирует также рассмотреть возможность применения этого аппарата к задачам анализа/синтеза звучащей речи.

Программное обеспечение системы реализовано в среде ОС Демос-86 (UNIX-совместимая ОС, прототип - ОС VENIX) и MS-DOS, на ПЭВМ типа PC с памятью 640 К (процессор типа Intel 8088) и жестким диском 20 М. Язык реализации - стандарт Си (K&R). Потребность в оперативной памяти не превышает 64 Кбайт, дисковой - ок. 2 Мбайт.

Литература

1. Добрушина Е.Р., Савина Г.Б., Гельбух А.Ф. Система точного морфологического анализа и синтеза // В сб.: Программное обеспечение новой информационной технологии. Калинин: АН СССР, НПО ЦПС, 1989.

2. Гельбух А.Ф. Разработка и программная реализация алгоритма точного морфологического анализа словоформ естественного языка на основе полного разделения лингвистической модели и программной оболочки // Дипломная работа в МГУ, мех-мат. факультет, 1990.