МОРФОАНАЛИЗ/СИНТЕЗ И ПРОВЕРКА РУССКИХ ТЕКСТОВ

А.Ф.Гельбух

Обработка текстов - одно из основных назначений ОС UNIX, оснащенной, кроме прекрасных форматтеров, таких, как nroff и troff, удобных многооконных редакторов, как red, мощных преобразователей текстов, от простого rpl до превосходного mm, еще и средствами работы с текстами на естественном языке, такими, как spell. Однако разработка средств для работы с текстами на русском языке, скажем, варианта того же spell, наталкивается на определенные трудности, связанные со сложностью грамматики русского языка по сравнению с английским.

Существует очень удобная (для своих применений) версия spell, хранящая в словаре все формы слов (*байт, байта, байту, байтами*...). Однако это, конечно, не решение проблемы, особенно когда дело доходит до глаголов и причастий.

Кроме того, для лингвистических процессоров, скажем, для синтаксического анализа, нужна настоящая полная морфология, и притом быстродействующая. Существует (и программно реализовано в от ОС ЕС до MS-DOS) много подходов к решению этой задачи (см. в [1]). Однако большинство как лингвистических моделей, так и реализаций опираются на неприемлемые для UNIX предполо жения об аппаратной или операционной среде, и прежде всего это - потребность в большой памяти для данных. Нет также и реализаций, которые, если бы их и можно было адаптировать к

UNIX, по своей идеологии органично вписались бы в его традиционную ткань. Настоящая разработка призвана восполнить этот пробел.

Она представляет собой оболочку системы морфологического анализа и синтеза текстов на любом флективном языке - то есть таком, в котором слова изменяются по суффиксам и окончаниям. К таким языкам относится русский и большинство языков мира. Структура конкретного языка хранится в таблицах (а лексика - в словаре) - текстовых файлах, полностью доступных администратору (пользователю) системы и открытых для любых дополнений, упрощений или изменений вплоть до создания таблиц для нового языка. Имеются специальные утилиты обслуживания таблиц и формирования словаря. Нет нужды и

говорить, что таблицы и словарь для русского языка имеются. Словарь построен на лексике классического словаря А.А.Зализняка; планируется пополнить его большим массивом технических терминов.

Анализ и синтез - полные и точные. Под точностью понимается соответствие результатов имеющимся в словаре и таблицах данным; не точным анализом мы назвали бы такой, который по слову *кровать* "догадался" бы, что это глагол, не имея или не найдя его, как некоторые промышленно эксплуатируемые на ЕС ЭВМ системы, в словаре. Под полнотой понимается рассмотрение всех омонимов и всех возможных вариантов разбора слова (*опала* - камня, *опала* - немилость, *опала* - теперь опавшая). Планируется дополнить систему обработкой ударений и буквы йо.

Для лиц, не владеющих русским языком, может оказаться очень полезной способность системы не только обьяснить, в каком падеже стоит заданное слово или какова его исходная форма (*шел* --> ИДТИ, мужской род, прошедшее время и т.д.), не только указать на ошибку в слове, но и предложить варианты его исправления (*краснИми* --> если КРАСНЫЙ, творительный падеж, то пишется КРАСНЫМИ. (хотя *синИми*)).

При разработке системы пришлось преодолеть две технические трудности: во-первых, фрагментированность файлов приводит к невозможности считывания с диска сразу, без дерганья головок дисковода, большого куска словаря. Значит, искать на диске основы слов "методом тыка"? И, во-вторых, необходимость экономить память. Решены они были так: не более одного обращения (считывание одного сектора) к диску на одно слово, что достигается оригинальной структурой словаря, и около 30 Кбайт на текст и код, вместе взятые. Быстродействие при анализе на машине IBM PC AT (12 Mhz) - около 30 слов/секунду, IBM PC базовой (под ОС Демос-86 - VENIX) - около 8 слов/секунду. Планируется увеличение быстродействия как минимум вдвое.

Система может быть использована двумя способами. Можно обращаться к отдельным ее функциям - анализ слова, синтез слова, исправление ошибки - из Си-программ как к библиотечным - этакая интеллектуальная <stdio.h>. Такой способ позволит строить особо быстродействующие прикладные программы. Например, синтаксический анализатор может, получив "нужный" омоним слова, приказать прекратить порождение других омонимов и перейти к следующему слову.

Другой вариант - использование программы анализа/синтеза в конвейере типа морфология | синтаксис | ИИ-система | синтаксис | морфология (слева - анализ, справа - синтез). Кроме того, средствами UNIX (csh, awk) ничего не стоит написать, например, маленький командный файл, который с помощью данной программы будет по входу "*фирме нужны новые машины. директор, важный, полученный, результат*" строить фразу "*директору важны полученные результаты*", что может оказаться полезным лицам, не владеющим русским языком. В общем, Ваша фантазия плюс неограниченные возможности UNIX могут помочь Вам успешно применить обсуждаемую программу от простейших электронных словарей до лингвистических процессоров и систем перевода.

1. *Гельбух А.Ф.* Простая оболочка системы точного морфологического анализа и синтеза текстов на естественном языке // В сб.: Использование программных средств ПЭВМ для автоматизации учрежденческой деятельности. Калинин, НПО ЦПС, 1990.