Библиотека для морфологического анализа свободнообразуемых слов русского языка.

Е. Е. Гончаренко*

*Факультет ВМК МГУ, gee.github@gmail.com

Введение

Задачей лингвистического процессора является преобразование естественно-языкового предложения (или даже целого текста) в некоторый набор семантических структур, являющихся формальным представлением "смысла" исходного предложения или текста. Цель такого преобразования — обеспечить исходные данные для работы поисковых механизмов СУБД. Вот список тех задач, в которых можно использовать лингвистический процессор:

1. написание переводчика;
2. задачи распознавания и синтеза речи;
3. распознавание текста;
4. проверка орфографии;
5. проверка синтаксиса;
6. информационно-поисковые системы;

Основным недостатком существующих лингвистических процессоров является чрезмерно большой объем словаря, порождающий ряд технических проблем:

- большие затраты труда на создание и поддержание словаря;
- невозможность полного размещения словаря в оперативной памяти компьютера при анализе;
- высокая избыточность информации, связанной с постоянными признаками каждой словоформы (морфологическими, синтаксическими, семантическими);

Современные компьютерные программы, анализирующие текст на естественном языке, как правило, используют словари. Цель словарей помочь распознать встреченную текстовую цепочку.

Целью данной работы является создание программы, которая, используя реально существующую лингвистическую базу данных, выдает морфологические характеристики некоторых слов, не содержащихся в этой базе данных. Программа основана на использовании морфологического анализа структуры незнакомого слова. Приблизительно анализ слова работает в такой последовательности. От предлагаемого слова отрезаются возможные префиксы, и оставшаяся часть проверяется на наличие в лингвистической базе данных. Если оставшаяся часть слова присутствует в базе данных, то в качестве информации об исходном слове, выдается полученная информация о части слова, с учетом всех префиксов.

1 Постановка задачи

Целью данной работы является создание программы, которая, используя реально существующую лингвистическую базу данных, выдает морфологические характеристики для следующих классов слов:

- свободнообразуемые слова;
- слова с дефисом;
- сложные слова.

Свободнообразуемые слова должны удовлетворять следующим условиям:

- Стандартность их соединения с существительными и прилагательными.
- Стандартность значения.
- Структурная самостоятельность.

Программа работает следующим образом:
От предлагаемого слова отрезаются возможные префиксы, и оставшаяся часть проверяется на наличие в лингвистической базе данных. Если оставшаяся часть слова присутствует в базе данных, то в качестве информации об исходном слове, выдается полученная информация (падеж, склонение и т.д.) о части слова, с учетом всех префиксов. Программа автоматически меняет основу, ударную букву, ставит второстепенное ударение.

2 Алгоритм анализа слова.

Анализ слова сводится к следующей последовательности действий:

1. На вход процедуры подается слово X.
2. Пытаемся найти в слове X дефис. Если мы его нашли, то ту часть слова X, которая была до дефиса (включая дефис), мы сохраняем как X1; а оставшуюся часть слова X как X2 и переходим к шагу 5, иначе на шаг 3.
3. Если слово X начинается на префикс из списка префиксов (см. пункт: Словарная информация для базы данных), то мы сохраняем этот префикс как X1, а оставшуюся часть слова X как X2 и переходим к шагу 5, иначе на шаг 4.
4. Если слово X начинается на порядковое числительное (например: тысячетрехсотдвадцатичетырехдневный), то мы сохраняем это числительное как X1, а оставшуюся часть слова X как X2 и переходим к шагу 5, иначе на шаг 8.
5. Обращаемся к морфологическому анализатору со словом X2. Если морфологический анализатор выдал морфологические характеристики слова X2, то перейти на шаг 6. Если же морфологический анализатор выдал, что слово не найдено, то перейти на шаг 7.

6. В качестве информации о слове X, выдается информация о слове X2, модифицированная следующим образом:
    - к основе слова X2 слева приписывается слово X1;
    - меняется номер ударной буквы;
    - ставится второстепенное ударение.
7. Рекурсивно вызываем данный алгоритм для слова X2. Если алгоритм выдал морфологические характеристики слова X2, то перейти на шаг 6. Если же алгоритм выдал, что слово не найдено, то перейти на шаг 8.
8. Слово не распознано. Стоп.