Элементарный лемматизатор

Нестеренко А.А.

Одна из наиболее сложных современных проблем в области компьютерных технологий – реализация общения с ЭВМ на естественном языке. В процессе решения поставленной задачи возникает множество подзадач, без решения которых это общение невозможно. В частности, трудность в обработке текстов на ЭВМ заключается в том, что машина воспринимает слово как набор символов и, следовательно, не может отождествить две различные грамматические формы одного и того же слова.

Когда человек читает текст, он автоматически ставит в соответствие основе каждого слова некоторое значение. Но тот процесс, который наш мозг может проделать за тысячные доли секунды, необходимо записать в виде алгоритма, который бы могла выполнить машина. Другими словами, программа должна провести лемматизацию. Значение алгоритмов лемматизации можно понять на примере такой системы, как электронный переводчик. Чтобы компьютер смог распознать все слова, встретившиеся в тексте, необходимо предоставить программе все возможные грамматические формы каждого слова. Это приводит к значительному увеличению объема базы данных. Выходом из этой ситуации может служить включение в систему алгоритмов лемматизации. Алгоритмы лемматизации позволяют объединить различные грамматические формы одного слова, эквивалентные по лексическому значению. А это объединение позволяет сделать следующий шаг к содержательной интерпретации текста.

Для осуществления лемматизации необходимо создание инструмента, в котором был бы реализован алгоритм, позволяющий при работе с текстом выделять основы и кортежи слов.

Поставленную задачу отчасти решает “Элементарный лемматизатор-2” – расширение существующей программы, разработанной студентом кафедры МО ЭВМ Плехановым А. Н. [1].

Основная идея алгоритма программы “Элементарный лемматизатор”– сравнение двух слов, начиная с первой буквы, и нахождение величины совпавшей части (рис. 1). Если величина общей части слов не ниже порогового значения, то рассматриваемые слова считаются родственными.

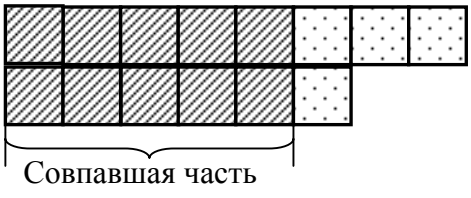

Рис.1. Сравнение слов в программе “Элементарный лемматизатор”

Под порогом будем понимать критерий, по которому определяется, являются ли две лексемы словоформами одного и того же слова. Критерием для сравнения слов может служить либо число совпавших букв, либо процент совпавших букв в этих словах. Порог подбирается экспериментально, и при некоторых значениях можно получить очень хорошие результаты. Для нахождения оптимального значения порога необходимо обработать словари большого объема.

Основная часть алгоритма лемматизации состоит в извлечении из текста слов, объединении их в группы родственных слов и, наконец, выделении в каждой группе совпадающих частей (основ) и несовпадающих (кортежей).

Разбиение исходного словаря на группы происходит следующим образом: программа последовательно берет в качестве входной информации слово из исходного словаря, которое ещё не отнесено ни к одной из полученных групп, и заносит его в группу, к словам которой оно подходит по порогу. Если же такой группы не обнаружено, формируется новая группа, состоящая только из текущего слова. Когда весь исходный словарь будет обработан, группы, состоящие из одного слова, объединяются в одну (нулевую) группу – одиночные слова.

Полученную информацию можно использовать в дальнейшем для формирования множеств основ с тождественными кортежами, отношений включения между ними, а также для формирования парадигм – множеств кортежей, с помощью установления отношений включения между кортежами. В парадигмы должны объединиться слова, имеющие одинаковые окончания в одинаковых формах (например, существительные одного склонения). Для тех слов, к которым в тексте не встретилось родственных, пользователь сам может указать окончание.

Планируется также реализовать блок, который будет, исходя из распределения частот встречи каждой формы одного слова в тексте, ставить в соответствие основе некоторое значение (например, обстоятельство места, обстоятельство времени, обозначение предмета и т.д.).

Этот блок позволит перейти к содержательной интерпретации текста, не отвлекаясь на множество грамматических форм слова, а также различать многозначные слова, если они используются в разных значениях (например, язык – орган, язык – средство общения).

Исходные данные для программы должны быть представлены в виде текстового файла. Полученные во время работы данные хранятся в таблицах базы данных. Также предоставлена возможность использовать ранее сохраненные таблицы в качестве входной информации для программы. Результаты работы приложения сохраняются в таблицах базы данных. Возможно сохранение исходного текста, в котором выделены окончания в словах, и словаря исходного текста.

Программа состоит из восьми функциональных частей:

1. Основной модуль – реализует большинство функций по обработке словаря: разбиение словаря на группы, выделение основ и кортежей в группах, поиск тождественных основ.
2. Открытие файла – модуль, содержащий процедуры обработки исходных файлов.
3. Открытие таблиц – модуль процедур, с помощью которых предоставляется возможность использовать ранее сохраненные таблицы.
4. Сохранение таблиц и словаря – модуль содержит процедуры сохранения данных.
5. Порог – реализация работы с порогом.
6. Иерархия кортежей – модуль с процедурами построения отношений включения между кортежами.
7. Одиночные слова – модуль содержит процедуры, позволяющие получить информацию об одиночных словах и выбрать для них окончания.
8. Формирование парадигм – модуль, содержащий процедуры формирования парадигм.

Взаимодействие модулей представлено на рис. 2.

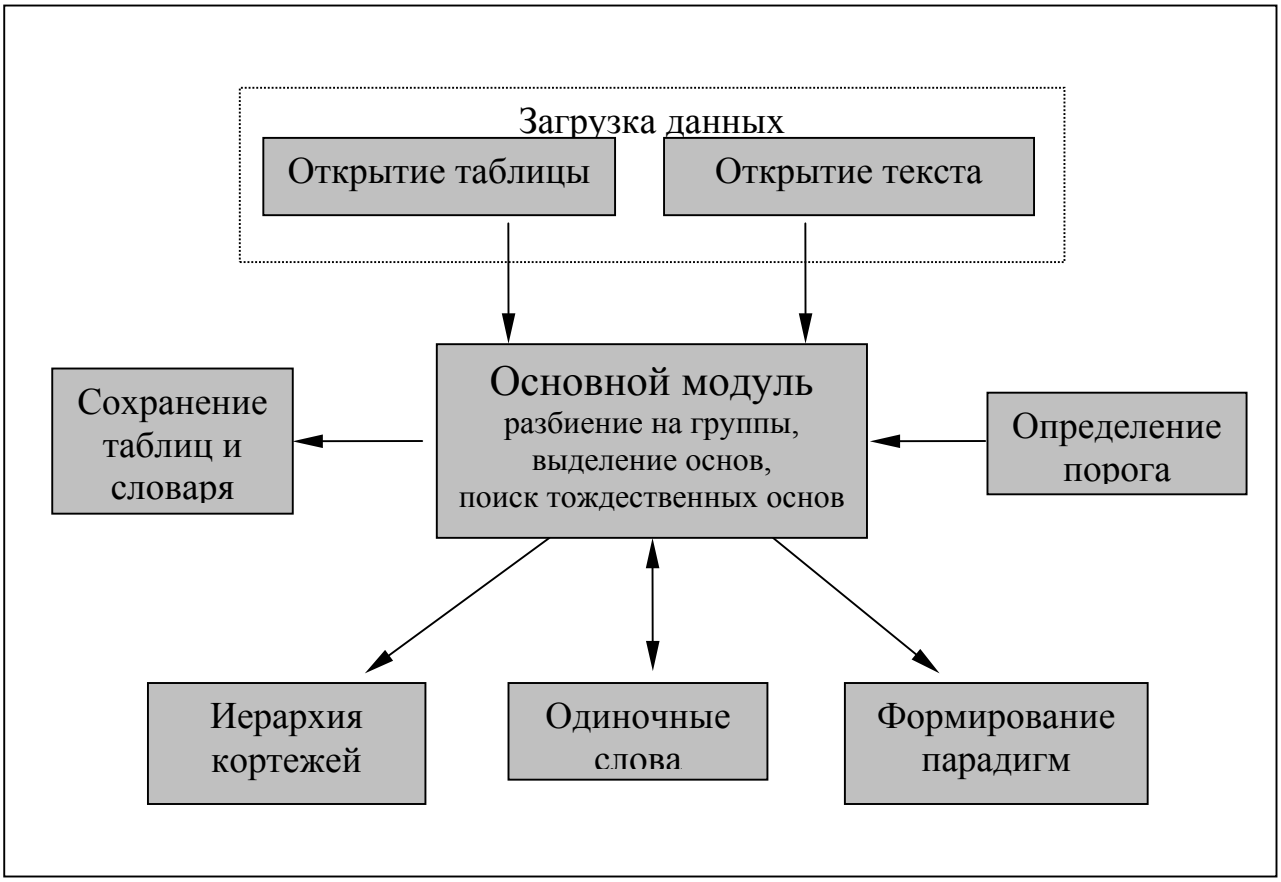

Рис. 2. Взаимодействие модулей программы.

Алгоритм, используемый в ”Элементарном лемматизаторе”, очень прост, и в этом его универсальность. Такой подход с одинаковым успехом может быть применен для отделения как окончаний, так и приставок, а далее суффиксов, что позволит перейти к корню слова. Кроме того, возможно применение программы для работы с различными языками. Немаловажно, что реализация всех этих возможностей не потребует значительных изменений в коде программы.

Литература:

1) Плеханов А.Н. Программно-инструментальные средства для содержательной интерпретации текста: Курсовая работа. Научно-методический центр компьютерной лингвистики. ВГУ. Воронеж. 2000.
2) Нахмансон Е.В. Программа лемматизации полных прилагательных, порядковых числительных, причастий и адъективных местоимений: Дипломная работа. Научно-методический центр компьютерной лингвистики. ВГУ. Воронеж. 1996.