Министерство образования и науки Российской Федерации
Государственное образовательное учреждение
Высшего профессионального образования
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»
им. В.И. Ульянова (Ленина)
СПбГЭТУ

Факультет компьютерных технологий и информатики
Кафедра автоматизированных систем обработки информации и управления

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к выпускной работе бакалавра на тему:

Реализация утилиты для работы со словарем АОТ

г. Санкт-Петербург
2012 г.

Оглавление

Введение

Системы морфологического анализа и синтеза развиваются уже не одно десятилетие, и серьёзная обработка текста уже, пожалуй, немыслима без их помощи. Как в России, так и за рубежом на рынке существуют много коммерческих программ, которые могут успешно справляться с этими задачами, но, к сожалению, они не могут быть использованы для научных экспериментов из-за их крайней высокой цены и отсутствия исходного кода. С другой стороны, существуют бесплатные модули, которые, впрочем, часто неприемлемы из-за низкой скорости обработки слов и неполноты словарных баз.

Морфологические словари с сайта aot.ru, которые я использовал, призваны решить указанную выше проблему, обеспечив научные коллективы и вообще любых возможных энтузиастов-экспериментаторов системой морфологического анализа и синтеза.

Глава 1

Морфологический словарь morphs.mrd

В данном дипломном проекте используется два морфологических словаря: русский и английский. Русский базируется на грамматическом словаре А.А. Зализняка 1987 г.[1] На данный момент он включает в себя 162519 лемм и 2553 наборов окончаний. Английский словарь создан на основе WordNet[2] и включает в себя 104657 лемм и 442 наборов окончаний. Словари имеют одинаковую структуру — они состоят из пяти разделов: наборы окончаний, наборы ударений, история изменений словаря, приставки, леммы. Каждая часть словаря начинается со строки, в которой указано количество строк в данном разделе, что дает возможность итератору вовремя остановиться. Изменяя файл, нужно следить за этими строками.

Первая часть словаря — это строки с наборами окончаний (флексии) вида:

«%ЫЙ*йа%ОГО*йб»,

где «ЫЙ», «ОГО» — окончания, а «йа», «йб» — анкоды, обозначающие граммемы леммы (интерпретацию нужно смотреть в rgramtab.tab или egramtab.tab).

Анкодом («аношкинским кодом») называется уникальный двухбуквенный идентификатор, который соответствует некоторой комбинации значений селективных признаков и граммем. Конечное множество аношкинских кодов исчисляет все встречающиеся в данном языке комбинации морфологических характеристик. Всего в морфологическом анализаторе русского языка системы Диалинг насчитывается 870 таких кодов.

Окончание также может быть и нулевое, например: «%*яд».

Каждая строка детерминирует отдельную парадигму, поэтому словооснова ссылается на номер строки, соответствующую нужной парадигме.

Следующим разделом идут наборы ударений, которые не используются в моей программе.

Затем идёт блок информации об истории внесения изменений создателями словаря, который также не используется в моей программе.

Далее идут приставки (префиксы), которые подставляются перед словоосновой.

Последним разделом является набор лемм вида:

«ЯХТСМЕН 51 43 1 Фб -»,

где «ЯХТСМЕН» — словооснова;

«51» — ссылка на набор окончаний (номер строки в разделе окончаний);

«43» — ссылка на набор ударений;

«1» — ссылка на набор приставок;

[1] Основополагающий труд по морфологии, где впервые был предложен системный подход к описанию грамматических парадигм, включающих не только изменение буквенного состава слов, но и ударения.

[2] Семантическая сеть для английского языка, разработанная в Принстонском университете, и выпущенная вместе с сопутствующим программным обеспечением под некопилефтной свободной лицензией.

«Фб» — ссылка на общие граммемы данной леммы (поле Ancode) (может быть «-»);

«-» — не реализовано на данный момент.

Общие граммемы данной леммы, это те граммемы, которые должны быть приписаны всем словоформам данной леммы, например, граммема «фам» (фамилия), или граммема «лок» (локативность). Это часто уже семантизированные граммемы. Набор приставок леммы — это те приставки, с которыми лемма образует полное слова языка. В набор приставок может входить пустая приставка, что означает, что лемма может быть использована сама по себе (без приставок).

Глава 2
Структура файлов словарей rgramtab.tab и egramtab.tab

Данный словарь служит для определения по анкоду найденного слова части речи, рода, склонения, падежа и т.п. Состоит он из пустых строк, комментариев (строк, начинающихся с «//») и строк вида: «аа А С мр,ед,им», где

«аа» — анкод;

«А» — ???;

«С» — код части речи (Все обозначения можно узнать в таблице 2.1);

«мр,ед,им» — набор граммем.

Таблица 2.1: Полный перечень русских частей речи.

Часть речи в системе Диалинг	Расшифровка	Пример
С	существительное	мама
П	прилагательное	красный
МС	местоимение-существительное	он
Г	глагол в личной форме	идет
ПРИЧАСТИЕ	причастие	идущий
ДЕЕПРИЧАСТИЕ	деепричастие	идя
ИНФИНИТИВ	инфинитив	идти
МС-ПРЕДК	местоимение-предикатив	нечего
МС-П	местоименное прилагательное	всякий
ЧИСЛ	числительное (количественное)	восемь
ЧИСЛ-П	порядковое числительное	восьмой
Н	наречие	круто
ПРЕДК	предикат	интересно
ПРЕДЛ	предлог	под
СОЮЗ	союз	и
МЕЖД	междометие	ой
ЧАСТ	частица	же, бы
ВВОДН	вводное слово	конечно
КР_ПРИЛ	краткое прилагательное	красива
КР_ПРИЧАСТИЕ	краткое причастие	построена

Граммема — это элементарный морфологический описатель, относящий словоформу к какому-то морфологическому классу, например, словоформе стол с леммой СТОЛ будут приписаны следующие наборы граммем: «мр, ед, им, но» и «мр, ед, вн, но». Таким образом, морфологический анализ выдает два варианта анализа словоформы стол с леммой СТОЛ внутри одной морфологической интерпретации: с винительным «вн» и именительным падежами «им».

Ниже перечислены все используемые граммемы:

мр, жр, ср — мужской, женский, средний род;

од, но — одушевленность, неодушевленность;

ед, мн — единственное, множественное число;

им, рд, дт, вн, тв, пр, зв — падежи, соответственно: именительный, родительный, дательный, винительный, творительный, предложный, звательный;

2 — обозначает второй родительный или второй предложный падежи;

св, нс — совершенный, несовершенный вид;

пе, нп — переходный, непереходный глагол;

дст, стр — действительный, страдательный залог;

нст, прш, буд — настоящее, прошедшее, будущее время;

пвл — повелительная форма глагола;

1л, 2л, 3л — первое, второе, третье лицо;

0 — неизменяемое;

кр — краткость (для прилагательных и причастий);

сравн — сравнительная форма (для прилагательных);

имя, фам, отч — имя, фамилия, отчество;

лок, орг — локативность, организация;

кач — качественное прилагательное;

вопр,относ — вопросительность и относительность (для наречий);

дфст — слово обычно не имеет множественного числа;

опч — частая опечатка или ошибка;

жарг, арх, проф — жаргонизм, архаизм, профессионализм;

аббр — аббревиатура;

безл — безличный глагол.

Как уже было сказано, одной словоформе может соответствовать много морфологических интерпретаций. Например, у словоформы СТАТЬ две интерпретации:

СТАТЬ, С, «но», («жр,ед,рд»,«жр,ед,дт», «жр,мн,им», «жр,мн,вн») ;

СТАТЬ, Г, «нп,св»,(«мн,дст,прш»).

Глава 3
Работа со словарем

Глава 4

Сравнение производительности

Работа программы, которая сначала загружает весь словарь в оперативную память и производит считывание и поиск в ней. Вывод программы, при подаче на вход книги Форда: Время чтения словаря: 4с 943мс Общее время: 10с 388мс Запросов к словарю: 59812. Успешно: 58457 Запросов к словарю в секунду: 10984

Работа программы, которая сначала загружает весь словарь в оперативную память, а затем по ходу выполнения создается кэш-файл также в оперативной памяти, производит считывание и поиск сначала в кэше, если в кэше данной записи еще не существует до ищет и дублирует её из словаря. Та же книга Форда: Время чтения словаря: 5с 660мс Общее время: 8с 547мс Запросов к словарю: 14455. Успешно: 13858 Cache hit: 45357. Cache miss: 14455 Запросов в секунду: 20717

Глава 5

Струкутра базы данных, хранящей словарь

rusDictionary create table flexiamodelsids (flexiamodelsid integer primary key); create table ancodes (ancode char(2) primary key, partOfSpeech varchar(15) not null, grammems varchar(35) not null); create table lemmata (baseStr varchar(30), flexiaModelId integer references flexiamodelsids (flexiamodelsid), counter integer, primary key (basestr, flexiamodelid)); create table flexiaModels (flexiaModelId integer references flexiamodelsids (flexiamodelsid), ancode char(2) references ancodes (ancode), flexiaStr varchar(20), counter integer, primary key (flexiamodelid, ancode, flexiastr));

Глава 6
Код