Алгоритм и практическая реализация морфемного разбора.

М. В. Едуш, П. В. Дикий

В работе рассматривается задача анализа текста на естественном языке. Рассматривается морфемный разбор слов текста как первый этап анализа. Приведён общий алгоритм разбора, baranbrus и краткое описание необходимых лингвистических правил. Проведено сравнение с существующими технологиями.

Введение

В настоящее время в лингвистике на первый план выходит задача построения различного рода лингвистических систем, которые выполняли бы роль конвертера человеческой речи в компьютерный формализованный язык. В связи с этим, одной из важных задач является разработка системы автоматического анализа текста. Одной из составляющих подобной системы является система морфологического разбора.

На сегодняшний день существует достаточно мало систем, проводящих автоматический разбор слова по составу. К тому же, эти системы используют словари основ и словоформ, а также используют базы с отношениями многие-ко-многим (M:M), что негативно сказывается на производительности БД, на объёме памяти для хранения информации, а значит и системы в целом. В отличии от них, разработанная система предполагает использование базы морфем и правил словообразования.

Описание принципа морфемного разбора

Морфемный разбор производится, опираясь на правила русского языка, но с учётом ограничений компьютера. В отличие от человеческого мозга, компьютер не может мыслить образами, следовательно ему необходим чёткий алгоритм разбора. Для этого определяем те части слова, которые можно определить абсолютно точно. Такими частями являются приставка и окончание, т.к. достоверно известно, что приставка находится в начале слова, окончание - самая последняя морфема. После этого остаётся выбрать корень или суффиксы, но специфика русского языка такова, что корень обычно является либо единственным, либо состоит из нескольких основ и соединительных гласных, таким образом, число возможных вариантов корней в слове значительно меньше числа вариантов расстановки суффиксов [4].

Исходя из этого, был выбран следующий порядок разбора:

1. поиск и определение возможных приставок;
2. поиск окончания;
3. поиск корней;
4. поиск суффиксов.

Было разработано два подхода к разбору слова:

- построение всех вариантов разбора с последующим отсечением противоречивых вариантов;
- нахождение произвольного разбора и уточнение его с помощью последовательного сдвига границ морфемных групп.

В обоих подходах применяется проход по слову с посимвольной проверкой вхождения текущего значения буферной строки с набором морфем данной части слова. Результатом проверки является одно из 3-х состояний: строка не имеет совпадений, строка является частью морфемы, строка является морфемой. На каждом этапе проверки идёт сужения диапазона проверяемых значений, что уменьшает общее время работы алгоритма. При полном совпадении морфема заносится в словарь. При потери совпадений начинаем поиск следующей морфемы.

Основным отличием подходов друг от друга является принцип выбора правильного варианта разбора после составления словаря морфем. В первом варианте выдаются все

возможные разборы слова (все возможные комбинации морфемных групп). Во втором варианте берётся некоторый разбор слова и модифицируется. Для этого выбирается максимально возможная последовательность приставок и корней. Далее происходит перебор суффиксов. Если длина суффиксной последовательности выходит за границу префикса и корня, то происходит перевыбор морфем для соответствия новой границе. При этом происходит поглощение вложенных суффиксов, приставок и корней.

В обоих вариантах производится проверка текущего разбора на соответствие орфографическим и лексическим правилам русского языка [5]:

- в слове не может быть двух одинаковых приставок или суффиксов за некоторыми исключениями. Примером может являться приставка "пра"(пра-пра-дедушка);
- морфемы, указывающие на разные части речи не могут находиться в одном слове (суффиксы "щий причастие, "ав деепричастие) и т.д.

Сравнение с существующими подходами

В качестве системы для сравнения была выбрана система морфемного разбора "ООМник"(http://www.oomnik.ru). Данная система также использует словари, однако в отличие от разработанной системы используются связи в базе типа многие-ко-многим, вместо набора морфем используется набор основ. Первоначальный поиск осуществляется по совпадению основ слова. Дальнейший поиск остальных частей осуществляется по совпадению.

Таким образом, можно вывести следующие особенности разработанного модуля:

- экономия дисковой памяти (хранение только морфем);
- возможность разбора любого слова, составленного по правилам естественного языка;
- независимость словарей различных типов морфем;
- привязка морфем к частям речи.

Выводы

Разработанный модуль является одним из ключевых элементов системы автоматизированного анализа текстов на естественном языке. Он позволяет разбирать слова русского языка, а также заимствованные слова по составу, опираясь на правила словообразования, орфографические и лексические правила. Набор правил и базы морфем легко расширяются за счёт отсутствия привязки их к конкретным морфемам. В реализации не используются сложные элементы, а, следовательно, и дальнейшая оптимизация не будет являться сложной. Модуль также способен определять часть речи входного слова, основываясь на той же базе правил.

Список литературы

[1] Черный А. И. Введение в теорию информационного поиска.- М.: Наука, 1975. - 238 с.

[2] Кент. А. Информационно-поисковые системы / Пер. с англ. - М., ВНИИУМ, 1965.

[3] Ланкастер Ф. Информационно-поисковые системы. Характеристики, испытания и оценка / Пер. с англ. - М.: Мир, 1972. - 308 с.

[4] М.Т.Баранов, Л.И.Григорян, Т.А.Ладыженская. “Русский язык” учебник для 7 класса общеобразовательных учреждений. — М.: Просвещение, 1997. — 191с.

[5] М.Т.Баранов, Л.И.Григорян, Т.А.Ладыженская. “Русский язык” учебник для 6 класса общеобразовательных учреждений. — М.: Просвещение, 1998. — 223с.

Авторы

Максим Владимирович Едуш — магистр 2-го года обучения, факультет автоматики и вычислительной техники, Таганрогский Технологический Институт Южного Федерального Университета, Таганрог, Россия; E-mail: *sh4d0w28@programmer.net*

Пётр Викторович Дикий — магистр 1-го года обучения, факультет автоматики и вычислительной техники, Таганрогский Технологический Институт Южного Федерального Университета, Таганрог, Россия; E-mail: *diki_petr@mail.ru*