УДК 004.43

Б.О. Кочконбаева, Н.Абдыраева
ст.преп. ОшТУ, преп. ОшТУ
B.O. Kochkonbaeva, N.Abdyraeva
Senior teacher OshTU, teacher OshTU

КОМПЬЮТЕРНАЯ ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

В данной статье рассматриваются вопросы искусственного интеллекта, в частности компьютерная обработка естественного языка. Рассматриваются вопросы создания алгоритма морфологического анализа текста.

Ключевые слова: интеллект, алгоритм текста, семантический анализ, структура.

COMPUTER PROCESSING OF NATURAL LANGUAGE

This article discusses the issues of artificial intelligence, such as computer processing of natural language. There are discussed the problems of creating an algorithm of morphological analysis of text.

Keywords: intelligence algorithm for text, semantic analysis, structure.

Компьютерная обработка естественного языка – область искусственного интеллекта, где решаются задачи общения человека с компьютером на естественном языке. Хотя задачи обработки текстов возникли практически сразу вслед за появлением вычислительной техники, но несмотря на полувековую историю исследований в области искусственного интеллекта, огромный скачок в развитии информационных технологий и смежных дисциплин, удовлетворительного решения большинства практических задач обработки текста пока нет.

Схема обработки текстов (рис.1) – независимо от того, на каком языке написан исходный текст, его анализ будет проходить все указанные стадии. Первые две стадии (разбиение текста на отдельные предложения и слова) практически одинаковы для большинства естественных языков. Единственное, где могут проявляться черты, специфичные для выбранного языка – это обработка сокращений слов и обработка знаков препинания (точнее, определение того, какие из знаков препинания являются концом предложения, а какие нет).

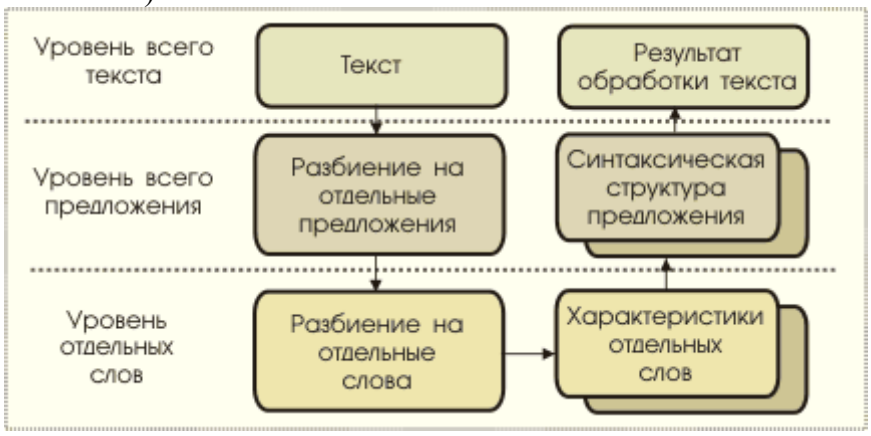

Рис 1. Общая схема обработки текста

Последующие две стадии (определение характеристик отдельных слов и синтаксический анализ), наоборот, очень сильно зависят от выбранного естественного

языка. Последняя стадия (семантический анализ), как и первые стадии, мало зависит от выбранного языка, но, это проявляется только в общих подходах к проведению анализа.

Существенную поддержку в проведении лингвистических исследований оказывают программы, позволяющие автоматически находить в исследуемых текстах нужные словоформы. Для этого должны быть составлены специальные программы, которые выполняют автоматический поиск словосочетаний.

Важной частью в автоматической обработке текстов на естественном языке является технология нахождения основы слова, родственный ей по целям алгоритм, позволяющий определить, что некоторая цепь словоформ составляют одну словоизменительную группу. Программа, способная выполнять эти операции включает в состав морфологический разбор слова в автоматическом режиме.

Проблема обработки текстов на кыргызском языке, «понимание» языка компьютером данное время является актуальной проблемой. Среди множества задач, которые сводятся к решению данной проблемы, можно назвать такие, как общение с компьютером на естественном языке (вопросно-ответные системы), информационный поиск, машинный перевод, извлечение полезной информации из текстов и т.д. Достаточно рутинная работа – проанализировать стилистику какого-либо автора по его работам. С помощью программы автоматического разбиения слов на морфемы и статистическим данным, появляется возможность автоматизированного анализа авторских текстов и составления готовых конкордансов.

Для этой цели было произведено изучение морфологии кыргызского языка. Правильное понимание состава слова, умение определить образующие его компоненты имеют большое значение при изучении языка. В слове отражены особенности строя языка, его лексико-семантические и функционально-грамматические законы.

Кыргызский язык отличается относительной регулярностью, позиционной и грамматической стабильностью морфологической структуры различных словоформ. Образование слов происходит последовательного присоединения к основе слова грамматических частиц – аффиксов (кыргыз+дар).

Представление структуры языка и систем анализа текста

Структурно-типологическая характеристика кыргызского языка связана с его принадлежностью к агглютинативным языкам. Для описания языков агглютинативного типа применяется набор признаков, учитывающих не только морфологические, но и синтаксические и фонетические особенности.

Морфологические признаки агглютинации:

1. Корень слова – в именительном падеже выступает в чистом виде, таким образом является центром всей парадигмы склонения;
2. Между морфемами четко сохраняется граница;
3. Строгая последовательность присоединения аффиксов;

Фонетические признаки агглютинации:

1. Наличие сингармонизма;
2. Фиксированное ударение, которое способствует сохранению фонетической целостности слова.

Синтаксические признаки агглютинации:

1. Твердый порядок слов в предложении;
2. Определение находится перед определяемым словом;
3. Дополняющее слово находится перед дополняемым словом;
4. Сказуемое в конце предложения.

Исходя из вышеизложенного нами было разработана алгоритм морфологического анализа естественного текста.

Алгоритм морфологического анализа

Алгоритм морфологического анализа по правилу в тексте заключается в следующем:

Модуль нормализации в процессе своей работы осуществляет следующую последовательность шагов:

1 шаг: Выполняется поиск слова в словаре начальных форм. Если слово в словаре найдено, то шаг 5.

2 шаг: Слово считывается посимвольно в обратном порядке (начиная с конца слова). Если слово закончилось, то работа алгоритма завершается. На основе текущего списка аффиксов формируется список гипотетических аффиксов.

3 шаг: Выполняется поиск всех гипотетический аффиксов в словаре аффиксов. Все найденные аффиксы добавляются в список аффиксов. Если ни один новый аффикс не найден, то переходим к шагу 2.

4 шаг: Выполняется поиск начальной части слова в словаре начальных форм. Если слово не найдено, то переходим к шагу 2.

5 шаг: В результат добавляется найденная основа и сопутствующий набор аффиксов.

Переход к шагу 2.

После нормализации, для каждого найденного слова осуществляется вычисление его морфологических характеристик на основе его аффиксов и морфологического класса основы.

После нормализации, для каждого найденного слова осуществляется вычисление его морфологических характеристик на основе его аффиксов и морфологического класса основы. Анализатор словоформ позволяет (Рис. 2.):

- Производить наполнение словаря начальных форм, если в словаре нет такой формы.
- При наполнении словаря у пользователя есть возможность удаления словоформы, если это не корректная форма слова.
- Модуль представляет возможность просмотра результа обработки слова.
- Отображается найденная в словаре основа слова, аффиксы которые были извлечены из словоформы.

Есть возможность загрузки словарей основ и аффиксов

В текстовое поле вводится слово для обработки при нажатии на кнопку анализ, начинается работа алгоритма по завершению, если такого слова нет в словаре, то происходит автоматическое добавление слова в словарь.

В противном случае в результате отображается найденная основа и список извлеченных аффиксов.

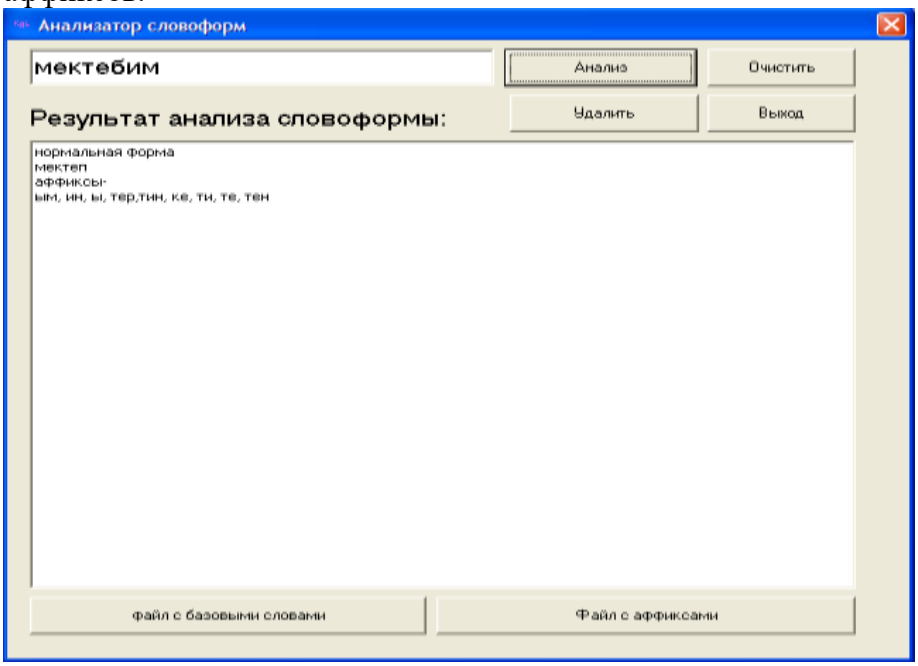

Рис.2 Интерфейс анализатора кыргызских словоформ

Выводы

Попытки формализовать интеллектуальную деятельность человека привели к постановке фундаментальной лингвистической задачи, состоящей в моделировании его языкового поведения, то есть в построении функциональной модели естественного языка.

За последнее десятилетие ярких результатов в области обработки текстов не было достигнуто, однако развитие новых технологий потребовало удовлетворительного решения некоторых задач обработки текстов. Развитие хранилищ данных делает актуальными задачи извлечения информации и формирования корректно построенных текстовых документов.

Обработка естественного языка или создание системы понимающей естественный язык состоит из определенных этапов. Первый этап обработки это создание морфологического анализатора. В процессе создания морфологического анализатора были догстигнуты следующие цели:

1. Изучение морфологии кыргызского языка, выделение морфологических классов
2. Построение морфологической таблицы для кыргызского языка
3. Реализован алгоритм анализа слов.

В дальнейшем планируется расширение пользовательского инфтерфейса и создание полнофункционального морфологического анализатора кыргызского языка.

Литература:

1. Азыркы кыргыз адабий тили: Фонетика, Лексикология, Лексикография, Фразеология, Морфология, Синтаксис стилистика, Текстаануу, Лингвопоэтика. // Бишкек -2009.
2. В.Е. Максимов, Л.А. Козленко, С.П. Маркин, И.А. Бойченко, Защищенная реляционная СУБД ЛИНТЕР. Открытые системы, 1999, №11-12
3. Орехов Б. В. Слободян Е. А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка [Текст] // Информационные технологии и письменное наследие: материалы ммеждународной научной конференции (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. — Уфа; Ижевск: Вагант, 2010. — С. 167–171.