УДК 81:004.912 http://doi.org/10.5281/zenodo.2290885

О МОРФОЛОГИЧЕСКОМ АНАЛИЗЕ В ПРИЛОЖЕНИЯХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА

©Кочконбаева Б. О., *Ошский технологический университет им. М. М. Адышева, г. Ош, Кыргызстан, buajar@mail.ru*

ABOUT MORPHOLOGICAL ANALYSIS IN NATURAL LANGUAGE PROCESSING APPLICATIONS

©Kochkonbaeva B., *Osh Technological University named after academician M. M. Adyshev, Osh, Kyrgyzstan, buajar@mail.ru*

Аннотация. Результаты морфологического исследования важны для всех прикладных приложений обработки естественного языка, которые важны для реализации процедуры разбора слов в тексте, прежде чем приступить к этапам семантико–синтаксического анализа. С этих позиций морфологический анализ должен считаться первым шагом на пути решения любой задачи компьютерной обработки естественного языка. В статье рассматриваются вопросы, связанные с использованием данных морфостатистики в различных приложениях обработки естественного языка, в частности, в создании и тестировании приложений автоматической обработки текста и их расширений на примере кыргызского языка.

Abstract. The results of morphological research are important for all applications of natural language processing that are important for the implementation of the word processing procedure in the text before proceeding with the stages of semantic–syntactic analysis. From this perspective, morphological analysis should be considered the first step in solving any problem of computer processing of a natural language. The article deals with issues related to the use of morfostatistics in various applications of natural language processing, in particular, in the creation and testing of natural language processing applications and their expansion using the example of the Kyrgyz language.

Ключевые слова: морфологичекий анализ, морфология, автоматическая обработка текста, АОТ, аффикс, морфема.

Keywords: morphological analysis, morphology, natural language processing, affix, morpheme.

Термин морфология заимствовано от греческого языка (morphe — форма, logos — наука), дает понятие «науки как о грамматической форме».

Итак, раздел морфологии исследует и изучает форму слов, словообразование и, составленную через форму слов, грамматическое значение [1]. Например, суу, суунун, сууга, сууну, сууда, суудан (вода, воды, воде, воду, от воды, в воде) и т. п. слова в морфологии являясь различной формой одного слова (здесь слова «вода»), определяется их смысловые особенности. Нижняя часть морфологического анализа определяется морфемой, верхняя — словом. Самая маленькая значимая единица называется морфемой. Морфология изучает систему слов, части речи и их категории.

Одним из основных понятий грамматики являются части речи. Иначе говоря, в кыргызском языке некоторые слова обозначают предмет, а иные — действие, число, признаки [2]. Поэтому слова в морфологии делятся на части речи.

Части речи кыргызского языка делятся на два: самостоятельные и служебные части речи. К самостоятельным частям речи входят: имя существительное, имя прилагательное, имя числительное, местоимение, глагол, наречие, а к служебным частям речи относятся — вводные слова, междометия, союзы, предлоги, частицы. Служебные части речи не могут дать никаких значений и не отвечают ни на какие вопросы. Например, жана (и), менен (с).

Автоматическая обработка языка является одной из наиболее быстро развивающихся областей в настоящее время. Для большинства языков, ресурсы необработанных текстов или даже простые словари — это ресурсы, которые крайне необходимы. Основанные на корпусах статистические исследования на естественном языке внесли новые измерения в лингвистическое описание и в различные приложения, разрешив некоторую степень автоматического анализа текста. Самый базовый формат, используемый при отображении информации о лингвистических элементах в корпусе, создается посредством перечисления и подсчета.

Одной из наиболее изученных областей автоматической обработки текста (АОТ) является морфология. Это важно, потому что язык продуктивен. В любом данном случае можно встретить текст, который содержит несколько слов и форм слов, которые мы не видели раньше, и которые не содержатся в любом предварительно скомпилированном словаре. Основная задача вычислительной морфологии состоит в том, чтобы взять слово в качестве входных данных и произвести для него морфологический анализ. Морфотактика определяет конкатенацию главных морфем слова и обычно описывается с помощью конечных автоматов. Но бывают ситуации, когда процесс формирования слов состоит не только в соединении морфем, но и в других процессах, таких как редупликация, инкорпорация и т. д. Это будут ситуации, когда фонологические правила призваны обеспечить обоснованность. Фонологические правила могут применяться и изменять форму морфов. Многие лингвисты моделировали фонологические правила, но считается, что наиболее успешной является модель, называемая двухуровневой морфологией. Двухуровневая морфологическая модель оказалась успешной для формализации морфологии очень разных языков (английский, немецкий и т. д.). Эта система используется даже для преобразования между различными системами записи.

Естественные языки состоят из очень большого количества слов, которые основаны на базовых строительных блоках, известных как морфемы, наименьшие лингвистические единицы, обладающие смыслом. В обработке естественного языка морфология используется для анализа внутренней структуры слов с использованием компьютеров. Результаты морфологического анализа текста имеет значение при работе таких прикладных программ, как морфологический анализатор, машинный перевод, экспертные системы, конечные автоматы и т. д.

Применение морфологичекого анализа естественного текста в прикладных программах рассматривается в следующих разделах.

Морфологический анализ — это процесс сегментации слов в морфемы или анализ процесса формирования слов. Это основной шаг для различных типов текстового анализа любого языка. Морфологический анализатор берет слово в качестве входных данных и вырабатывает корень, а его грамматические функции — как результат. Например, на основе морфологичекого анализа в среде RAD Studio была составлена программа анализатора

«NLP». Рассмотрим работу системы: приложение состоит из базы данных и интерфейса пользователя.

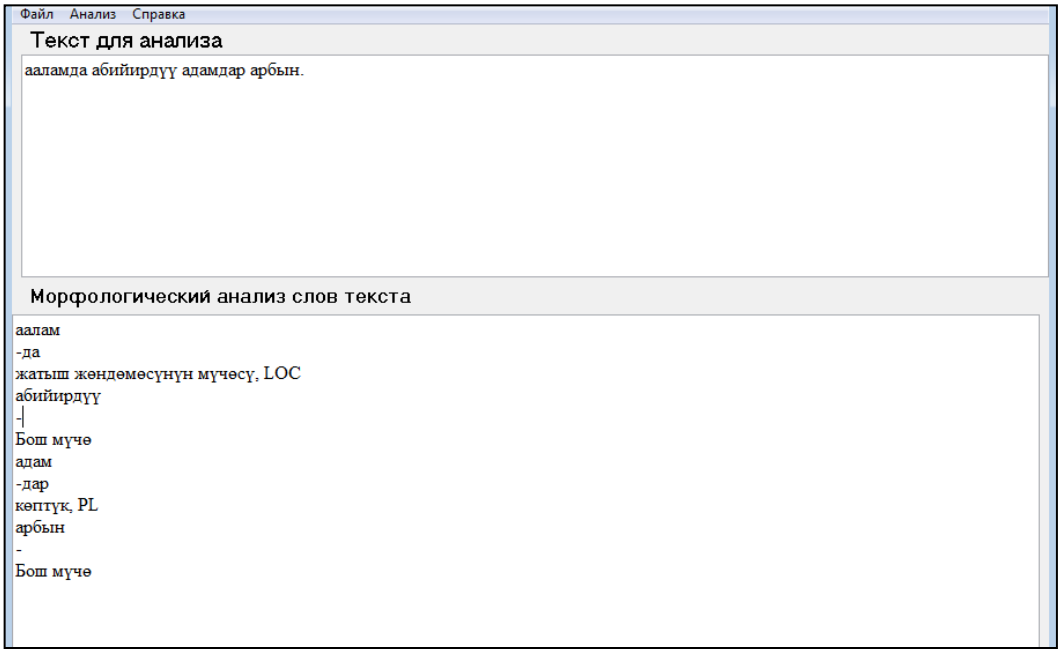

Рисунок 1. Результат работы системы «NLP».

Морфологический анализ — очень важный шаг к эффективному АОТ для высокоинформационных языков. Морфология является одной из дополнительных частей структурных аспектов выражения естественного языка. Это исследование особенно важно, потому что, помимо приобретения морфологии, мы могли бы сосредоточиться на генерации естественного языка, а также на возможных разложениях данного слова.

Машинный перевод

Рассмотрим алгоритм машинного перевода, основанного на лингвистическом анализе [4].

Шаг 1. *Получение предложения исходного текста из файла или из буфера в памяти.*

Шаг 2. *Разбиение предложения на слова и определение границ предложения.*

Шаг 3. *Морфологический анализ исходного текста — получение всех возможных лексических кодов для каждого найденного в словаре слова.*

Шаг 4. *Синтаксический анализ исходного текста — группировка однородных прилагательных и существительных, построение дерева главных/зависимых слов.*

Шаг 5. *Семантический анализ исходного текста.*

Шаг 6. *Осуществление перевода построенного дерева.*

Шаг 7. *Осуществление согласования переведенного дерева — семантический, синтаксический и морфологический синтез.*

Шаг 8. *Запись переведенного предложения в файл или в буфер.*

Из алгоритма видно, что в машинном переводе основную роль играет морфологический анализ и далее о работе этого модуля.

Решение данной задачи базируется на словаре исходного языка. В результате поиска по словарю каждому слову предложения приписывается множество лексико–грамматических классов: часть речи, падеж, число, категория и т. д., что позволяет в дальнейшем производить

сравнение классов, основанное на определенных характеристиках (например, проверять согласование прилагательных и существительных). Процесс поиска слов по словарю предполагает, кроме поиска оригинального слова в случае, если оно не было найдено в словаре, поиск слов с удалением возможных аффиксов. Для эффективного поиска аффиксов используется древовидная структура, элементами которой являются буквы аффиксов. Поиск останавливается либо когда нет дальнейшего перехода в дереве, либо когда найден аффикс и слово без этого аффикса существует в словаре. Кроме словаря аффиксов, для каждого из языков существует таблица межъязыкового соответствия, с помощью которой на этапе синтеза текста получается результирующее слово. На этапе распознавания классов производится также выделение словосочетаний, которые, согласно словарю, переводятся одним словом. Далее считается, что все такие словосочетания представляются одним словом. Это гарантирует правильность согласования и перевода словосочетания как единого целого.

Как мы уже упоминали ранее, морфологическое исследование является важной задачей, разделяемой многими приложениями в области АОТ. Тесно связанная с этой задачей одна из основных проблем при разработке таких приложений заключается в том, как организовать и сохранить в лексике морфологическую информацию, необходимую для анализа и генерации слов. Машинный перевод, как правило, является одним из приложений, которые одновременно обрабатывают процессы анализа и генерации. Лингвистические базы данных для систем МП должны быть разработаны таким образом, чтобы знания, которые они хранят, были максимально независимы от процесса. Таким образом, создание декларативных баз данных лексики в приложениях МП является обязательным.

Применение в речевых технологиях

Модели вариаций произношения непосредственно применимы к некоторым отраслям речевых технологий, таким как системы преобразования текста в речь (ПТР) и системы автоматического распознавания речи (АРР). Система ПТР, как следует из названия, преобразует письменный текст в речь. Система имеет множество компонентов, таких как анализ текста, прогнозирование продолжительности и предсказание интонации. Вся соответствующая информация отправляется компоненту синтеза речи для создания речевого вывода.

Компонент текстового анализа имитирует способность образованного читателя. Письменный текст анализируется и преобразуется в представление с морфологической информацией, структурами слогов, просодическими тегами, строками фонем и строками морфемы. Прогнозирование вариации произношения — одна из областей, которая представляет интересные возможности. Это связано прежде всего с тем, что системы все еще ограничены режимом чтения.

Правильное обращение с вариантами произношения будет полезно для создания правдоподобной речи, а также для моделирования личных характеристик и региональных акцентов. Распознавание речи (АРР) — это обратный процесс ПТР. Он принимает речь в качестве входных данных и преобразует ее в текст. ПТР является автоматическим считывателем, а АРР является автоматическим транскриптором. Моделирование произношения является важнейшим компонентом системы АРР. Таким оразом, морфологические анализы играют важную роль в системах ПТР и АРР.

Заключение

Исследование морфологии кыргызского языка значительно улучшилось в последние десятилетия. Эти улучшения оказали положительное влияние на области вычислительной

лингвистики. Многие приложения АОТ, такие как морфологический анализатор, морфологический генератор, машинный перевод, проверка орфографии и т. д. проходят через модуль морфологического анализатора.

Исследования в области АОТ улучшаются с каждым годом и хорошие морфологические анализаторы являются основой лучшего машинного перевода и робототехники.

Список литературы:

1. Абдувалиев И., Садыков Т. Азыркы кыргыз тили [Морфология]. Бишкек, 1997. (на кирг. яз).
2. Батманов И. А. Части речи в киргизском языке. Фрунзе, 1936.
3. Белоногов Г. Г., Богатырев В. И. Автоматизированные информационные системы / под ред. К. В. Тараканова. М.: Сов. радио, 1973. 328 с.
4. Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. 256 с.
5. Воронович В. В. Машинный перевод. Конспект лекций. Минск, 2013.

References:

1. Abduvaliev, I., & Sadykov, T. (1997). Azyrky kyrgyz tili [Morphology]. Bishkek. (in Kyrgyz).
2. Batmanov, I. A. (1936). Chasti rechi v kirgizskom yazyke. Frunze. (in Russian).
3. Belonogov, G. G., & Bogatyrev, V. I. (1973). Avtomatizirovannye informatsionnye sistemy. Ed. by K. V. Tarakanov. Moscow, Sov. radio, 328. (in Russian).
4. Apresyan, Yu. D., Boguslavskii, I. M., Iomdin, L. L., & al. (1992). Lingvisticheskii protsessor dlya slozhnykh informatsionnykh sistem. Moscow, Nauka, 256. (in Russian).
5. Voronovich, V. V. (2013). Mashinnyi perevod. Konspekt lektsii. Minsk, 2013. (in Russian).

Работа поступила
в редакцию 15.11.2018 г.

Принята к публикации
19.11.2018 г.

Ссылка для цитирования:
Кочконбаева Б. О. О морфологическом анализе в приложениях автоматической обработки текста // Бюллетень науки и практики. 2018. Т. 4. №12. С. 608-612. Режим доступа: http://www.bulletennauki.com/12-32 (дата обращения 15.12.2018).

Cite as (APA):
Kochkonbaeva, B. (2018). About morphological analysis in natural language processing applications. *Bulletin of Science and Practice, 4*(12), 608-612. (in Russian).