УДК 519.7

Р.Р.Мусабаев, К.Ч.Койбагаров, А.Т.Абдрахманов

Институт проблем информатики и управления, Алматы (E-mail: ata61@mail.ru)

Морфологический анализ текстов на казахском языке

В статье рассмотрены особенности алгоритмической реализации морфологического анализа текстов на казахском языке с учётом его структуры. Изучены особенности реализации технологии графематического анализа текстов на естественном языке и морфологического анализа казахских словоформ. Приведена последовательность применения правил морфологического разбора словоформ. Предложен обзор применения данной технологии к языкам, принадлежащим к классу агглютинативных языков.

Ключевые слова: морфологический анализатор, словоформа, аффикс, лексема, синтаксический анализ, семантический анализ, графематический анализ, токен, парсер, лемматизация, смысловой поиск.

1. Введение

Использование современных информационных технологий позволяет интенсифицировать применение казахского языка в компьютерных системах, таких как настольные, *web* и мобильные приложения, а также в облачных технологиях. Внедрение технологий грамматического анализа текстов на казахском языке способствует улучшению качества информации на данном языке в глобальных информационных ресурсах. Также повышается удобство и возрастает скорость применения языка в современных компьютерных приложениях.

Казахский язык принадлежит к тюркской семье языков, куда относятся также узбекский, киргизский, татарский, башкирский, азербайджанский, турецкий и др. Казахский язык относится к классу агглютинативных языков. Для этого класса языков характерно присоединение однозначных суффиксов или окончаний, несущих грамматическое значение, к неизменяемому корню или основе, являющихся носителями лексического значения.

Агглютинативный строй казахского языка намного облегчает задачу формирования словоформ по известной основе слова, так как любое словоизменение образуется посредством последовательного присоединения к основе слова соответствующих аффиксов, а также путем присоединения в виде цепочек одних аффиксов к другим.

Таким образом, для корневых или производных слов-основ можно формальным путем воссоздать всевозможные аффиксальные словоизменения — словоформы по известным грамматическим правилам [1].

Нам необходима обратная задача по разбору слова из текста на основы слова и присоединенных аффиксов.

2. Технология морфологического анализатора

Для решения данной задачи нами разработаны программы на ObjectPascal с использованием инструментальной среды разработки Delphi 7. В качестве системы управления базами данных используется MSSQL 2008. Доступ к данным осуществляется с помощью языка структурированных запросов (SQL) с применением технологии dbExpress.

В качестве исходных лексических материалов используется собранный нами морфологический словарь из 43 тыс. слов-основ. Словарь основных аффиксов насчитывает 240 аффиксов, которые служат кирпичиками для составления более сложных слов. Ведутся работы по составлению словарей имен и географических названий, а также словаря фразеологизмов. Кроме того, ведутся работы по составлению Национального корпуса текстов на казахском языке, и к настоящему времени собранный материал составляет около 200 мегабайт, насчитывающий 500 тыс. слов различных жанров и направлений.

В морфологическом словаре слова представлены в единственном числе и именительном падеже. Слово из единственного числа и именительного падежа преобразуется во множественное число и другие падежи путем присоединения аффиксов по определенным правилам.

Таким образом, задача по определению, в каком числе и падеже находится слово, состоит в выделении из слова леммы и аффикса. Затем аффикс разбирается на составные аффиксы. По словарю основных аффиксов мы сможем определить число и падеж данного слова из текста (рис. 1).

Рисунок 1. Схема морфологического анализа текстов на казахском языке

В результате разработки морфологического анализатора мы сможем получить следующие свойства:

– учет опечаток и выбор верного исправления в процессе анализа;
– возможность получения частичного разбора в случае ошибок в тексте;
– учет в процессе анализа как лингвистической составляющей (возможность и невозможность существования определённых конструкций), так и статистической (оценивание вероятностей конструкций для построения решения с максимальным правдоподобием).

Компоненты, составляющие языковую модель, — лингвистические процессоры, которые друг за другом обрабатывают входной текст (рис. 2). Вход одного процессора является выходом другого. Выделяются следующие компоненты [2]:

– графематический анализ. Выделение слов, цифровых комплексов, формул и т.д.;
– морфологический анализ. Построение морфологической интерпретации слов входного текста;
– синтаксический анализ. Построение дерева зависимостей всего предложения;
– семантический анализ. Построение семантического графа текста.

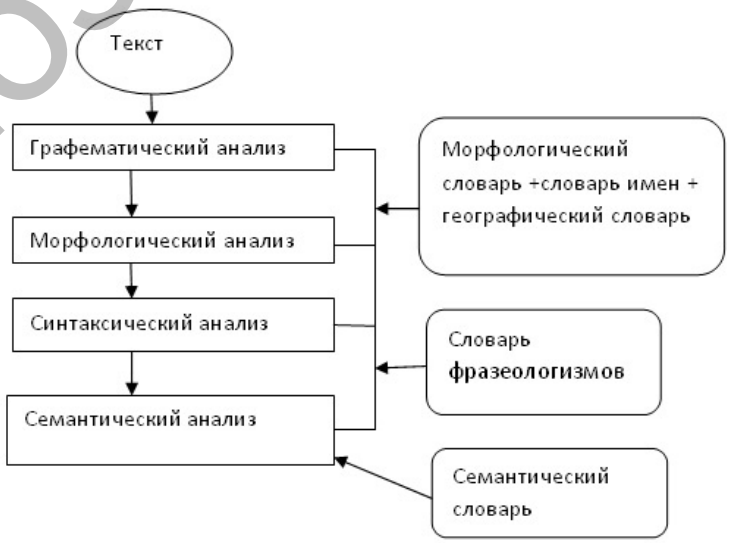

Рисунок 2. Компоненты языковой модели (лингвистические процессоры)

Пока нами решается задача по разработке морфологического анализатора, в который будет входить графематический анализатор (парсер).

На вход парсера подается текст на казахском языке. После выделения токенов текст подвергается морфологическому анализу. На каждом этапе цепочка токенов подвергается обработке системы продукционных правил.

Первая группа правил определяет позицию токена в абзаце или предложении. Вторая группа правил проверяет тип токена: знак пунктуации, наличие цифры, регистр. Следующая группа анализирует морфологические характеристики слова, поскольку одному токену могут соответствовать несколько разных лексем, а каждой лексеме — различные грамматические параметры (падеж, число, время и др.).

3. Графематический анализ

В задачу графематического анализа входят:

1. Разделение входного текста на слова, разделители и т.д.
2. Выделение устойчивых оборотов, не имеющих словоизменительных вариантов.
3. Выделение Ф.И.О (фамилия, имя, отчество), когда имя и отчество написаны инициалами.
4. Выделение электронных адресов и имен файлов.
5. Выделение предложений из входного текста.
6. Выделение абзацев, заголовков, примечаний.

Выделение токенов

После считывания очередного абзаца текста графематический анализатор разбирает токены и приписывает им соответствующие графематические характеристики. На этом этапе выделение токенов производится по пробелам и знакам препинания. Однако в ряде случаев лексические единицы в предложении имеют более сложную структуру. Первая группа правил и производит синтез таких сложных токенов, это могут быть, например, сложные единицы измерения (*кв.м*, *км/час*), интернет-адреса (*http://yandex.ru*), записанные цифрами порядковые числительные (*1917 ж.*), выделения фамилий (*Қасымов К.С.*).

Одновременно по Словарю фразеологизмов объединяются в один токен неизменяемые слова, образующие сложные наречия, существительные и т.д. Затем производится выявление инициалов и аббревиатур и подключение их к словам-хозяевам, например, имена и отчества — к фамилиям, буква, обозначающая слово год (ж.) — к дате и т.д.

В результате становится ясно, что некоторые точки не являются границами предложений, несмотря на то, что после них стоит прописная буква. Только после этого выполняется разбивка абзаца на предложения, и весь дальнейший анализ ведется уже только в пределах одного предложения.

4. Морфологический анализ

Морфологический анализатор базируется на Словаре основ-слов, состоящем из 43 тыс. слов (лемм), Словаре имен и фамилий, Словаре географических слов, а также Словаре базовых аффиксов, состоящем на данный момент из 240 аффиксов.

Морфологический компонент осуществляет морфоанализ и лемматизацию казахских словоформ (лемматизация — приведение текстовых форм слова к словарным; морфоанализ — приписывание словоформам морфологической информации). Лемма — это нормальная форма слова. Например, для существительных — это единственное число (если оно есть у существительного), именительный падеж.

На вход морфологического анализатора поступает цепочка токенов от предыдущего анализатора. Далее следует процедура лемматизации, из цепочки токенов выбираются токены, содержащие только буквы. Если входная словоформа не была найдена в словаре, то используется алгоритм предсказания, который ищет в словаре словоформу, максимально совпадающую с конца со входной словоформой. Найденная лемма будет служить основой словоформы. Оставшуюся часть словоформы будем считать аффиксом, который послужит для дальнейшего разбора на составные аффиксы; с использованием правил присоединения аффиксов друг к другу.

Как говорилось выше, Морфологический словарь состоит из слов в единственном числе и именительном падеже, глаголы представляют собой форму 2-го лица единственного числа повелительного наклонения, а аффиксы определяют число и падеж словоформы. Таким образом, задача определения морфологической характеристики каждой лексемы особых проблем не вызывает (рис. 1). Боль-

шую сложность вызывает задача по определению правильности соединения аффиксов друг к другу. Для решения этой задачи мы сгруппировали аффиксы для существительных и глаголов. Однако есть некоторое число аффиксов, которые могут использоваться и в существительных, и в глаголах.

Данная работа осуществляется в рамках реализации проекта «Разработка системы смыслового поиска текстов нового поколения, ориентированной на казахский язык». Также результаты данной работы предполагается использовать при реализации систем синтеза и распознавания речевого сигнала на казахском языке, в системах грамматического анализа в составе офисных пакетов программ, а также в системах оптического распознавания текстов (OCR).

References

1 *Zhubanov A.K.* Basic principles of formalization of the contents of the Kazakh text // Dis. Doc. — Almaty: Zhazushy, 2002. — P. 309.

2 *Sokirko A.V*. Morphological modules on a site www.aot.ru // Dialogue', 2004. June, 2–7.

Р.Р.Мұсабаев, Қ.Ч.Қойбағаров, А.Т.Әбдірахманов

Мәтіндердің қазақ тілінде морфологиялық талдауы

Мақалада мәтіндердің қазақ тілінде, оның құрылымын есепке ала отырып, алгоритмді іске асыру жолымен морфологиялық талдауын жасау ерекшелігі зерттелген. Мәтіндердің табиғи тілде графематикалық талдау технологиясының іске асыруының ерекшелігі, сонымен қатар қазақ сөз түрлерінің морфологиялық талдауы қарастырылған. Сөз түрлерінің морфологиялық талдауының ережелерінің қолдану тізбегі келтірілген. Аталған технологияның агглютинативті тілдер класына қатысты тілдерге қолданылуына шолу жасалған.

R.R.Musabayev, K.Ch.Koibagarov, A.T.Abdrakhmanov

The morphological analysis of texts in the kazakh language

In this work features of algorithmic implementation of the morphological analysis of texts in the kazakh language taking into account its structure are considered. Features of realization of technology of the grafematical analysis of texts in a natural language and the morphological analysis of the Kazakh word forms are considered. The sequence of application of rules of morphological analysis of word forms is given. The review of application of this technology with reference to the languages belonging to a class of agglutinative languages is provided.