УДК 004.421: 81-114.2 https://doi.org/10.33619/2414-2948/40/28

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗА КЫРГЫЗСКОГО ЯЗЫКА

©Сатыбаев А. Д., *д-р физ.-мат. наук, Ошский технологический университет им. М. М. Адышева, г. Ош, Кыргызстан, abdu—satybaev@mail.ru*
©Кочконбаева Б. О., *Ошский технологический университет им. М. М. Адышева, г. Ош, Кыргызстан, buajar@mail.ru*

MATHEMATICAL MODELING AND ALGORITHM OF MORPHOLOGICAL ANALYSIS OF THE KYRGYZ LANGUAGE

©Satybaev A., *Dr. habil, Osh Technological University named by M.M. Adyshev, Osh, Kyrgyzstan, abdu-satybaev@mail.ru*
©Kochkonbaeva B., *Osh Technological University named by M. M. Adyshev, Osh, Kyrgyzstan, buajar@mail.ru*

Аннотация. Исследования морфологического анализа языка дает нам дальнейшую обработку языка, так как морфологический анализ считается первым шагом на пути решения любой задачи компьютерной обработки естественного языка. В статье рассматриваются вопросы создания математической модели кыргызского языка и алгоритмы морфологического анализатора.

Abstract. Studies of the morphological analysis of the language gives us further processing of the language, since morphological analysis is considered the first step towards solving any problem of computer processing of natural language. The article deals with the issues of creating a mathematical model of the Kyrgyz language and algorithms of a morphological analyzer.

Ключевые слова: функция, естественный язык, словарь, морфология, кыргызский язык, алгоритм, словоформа.

Keywords: function, natural language, dictionary, morphology, Kyrgyz language, algorithm, word form.

Введение

Морфологический анализ является начальной ступенью различных задач, связанных с естественным языком, и поэтому его точное выполнение имеет большое значение.

Методы морфологического анализа можно разделить на 3 типа:

– анализировать со словарем аффиксов;

– анализировать с помощью словаря аффиксов и основ;

– анализировать с помощью словаря системы слов.

В методе анализа с помощью словаря аффиксов рассматривается выделение аффиксов из слова и поиск по словарю, и на этой основе раскрыть грамматическое значение слова.

Математическое моделирование

Обозначим словоформу в любом агглютинативном языке строкой $S_n = x_1 x_2 \dots x_n$, где x_i *(i=1, 2,…,n)* является членом соответствующего алфавита А, а *n* является количеством букв

(то есть длиной строки). В исследовании используем кыргызский алфавит, который состоит из 36 букв и знака подчеркивания _ для пустого символа следующим образом:

A={а,б,в,г,д,е,ё,ж,з,и,й,к,л,м,н,ң,о,ө,п,р,с,т,у,ү,ф,х,ц,ч,ш,щ,ъ,ы,ь,э,ю,я,'_'}

и мы ввели следующие обозначения S_n для обозначения подстрок любой строки $1 \leq i \leq j \leq n$:

$$S_n[i:j] = x_i x_{i+1} \dots x_j$$
$$S_n[:j] = x_1 x_2 \dots x_j$$
$$S_n[i:] = x_i x_{i+1} \dots x_n$$

Исходя из наших обозначений, специальная подстрока $S_n[i:i+1] = x_i x_{i+1}$ обозначается упорядоченной парой букв $(x_1,x_2)_i$, где субиндекс i $(i=1,2,\dots,n-1)$, указывает начальную позицию упорядоченной пары в этой строке $x_1=x_i$, $x_2=x_{i+1} \in A$.

Для $i=n$ упорядоченная пара формируется добавленным пробелом как $(x_n,'\_')_{i=n}$. Таким образом, любая строка $S_n = x_1 x_2 \dots x_n$ имеет n упорядоченную пару в нашем исследовании.

Для заданной упорядоченной пары букв $(x_1,x_2)_j$ которая может появляться в позиции $1 \leq j \leq n_{max}$ в любой форме кыргызского слова (где n_{max} максимальная длина слова в кыргызском языке) и данная конкретная форма слова обозначается как $S_n = x_1 x_2 \dots x_n$, где $n \geq j$, , обозначение $(x_1,x_2)_j \in S_n$ указывает, что существует упорядоченная пара $(x_1,x_2)_i$ в позиции i $(1 \leq i \leq n)$ в S_n при условии, что $(x_1,x_2)_i = (x_1,x_2)_j$ *для* $i=j$. Наконец, мы определяем еще два символа, а именно $g_m=S_n[:m]$ *и* $e_m=S_n[m:]$ чтобы представить любую словесную форму в виде упорядоченной пары из двух подстрок $S_n^m = (g_m, e_m)$ для всех $1 \leq m \leq n$.

Предположим, что множество L будет набором всех возможных упорядоченных пар букв $(x_1,x_2)_i$ которое может появляться в любой кыргызской словесной форме для позиций $i=1,\dots n_{max}$. Тогда L будет пробным пространством и может быть определено следующим образом:

$$L = \{(x_1,x_2)_i | x_1, x_2 \in A \; and \; 1 \leq i \leq n_{max}\}$$

И далее предположим, что множества G_k , E_k и T_k, *где* G_k , E_k , $T_k \subset L$, $1 \leq k \leq n_{max}$ представляют события, определенные следующим образом:

$$G_k = \{(x_1,x_2)_i | i = k \; and \; (x_1,x_2)_i \in g_m \; and \; 1 \leq m \leq n_{max}\}$$
$$E_k = \{(x_1,x_2)_i | i = k \; and \; (x_1,x_2)_i \in e_m \; and \; 1 \leq m \leq n_{max}\}$$
$$T_k = \{(x_1,x_2)_i | i = k, \; h_1 = s_n[k:k], \; h_2 = s_n[k+1,k+1], \; 1 \leq i \leq n_{max}\}$$

Таким образом, для каждой упорядоченной пары $(x_1,x_2)_i$ в позициях $i=1,2,\dots,n$ любой заданной словоформы, обозначенной через $S_n = x_1 x_2 \dots x_n$ можно определить вероятности нахождения в вышеуказанных три множества следующим образом:

$$Pr(s_n[i:i+1] \in G_i) = Pr((x_1,x_2)_i \in G_i) = P_G((x_1,x_2)_i) \quad (1)$$

$$Pr(s_n[i:i+1] \in E_i) = Pr((x_1,x_2)_i \in E_i) = P_E((x_1,x_2)_i) \quad (2)$$

$$Pr(s_n[i:i+1] \in T_i) = Pr((x_1,x_2)_i \in T_i) = P_T((x_1,x_2)_i) \quad (3)$$

Где, уравнение (1) относится к вероятности того, что упорядоченная пара $(x_1,x_2)_i$ находится в основной части заданной формы слова, аналогично уравнение (2) относится к вероятности того, что упорядоченная пара $(x_1,x_2)_i$ находится в аффиксной части данной

формы слова и, наконец, уравнение (3) относится к вероятности того, что упорядоченная пара $(x_1, x_2)_i$ находится между частью основы и аффиксной частью данной формы слова (то есть, x_1 — последняя буква части стебля, а x_2 — первая буква части аффикса).

Ввиду того, что слова кыргызского языка состоят из корня и аффиксов, слово обозначим как S, тогда в качестве функции их определим так:

$$S = R + \sum_{i=0}^{m} U_i \,, (m \leq 8) \tag{4}$$

Здесь, *S* — линейная функция, *R* — основа слова, *Um* — словоизменительные аффиксы.

В соответствии с формулой (4) *S* зависит от корня, словообразовательных аффиксов, словоизменительных аффиксов.

Словоизменительные аффиксы могут достичь до восьми, иначе говоря

$$\sum_{i=0}^{8} U_i = U_0 + U_1 + U_2 + \cdots + U_8, \tag{5}$$

Определение 1: Если $Km=\varnothing$, $Um=\varnothing$, то *S* функция будет равна корню слова, и вводимое слово не разделится на морфемы.

Множество словоизменительных аффиксов

Словоизменительные аффиксы изменяют грамматическое значение слов, но не изменяют лексическое значение.

Группируя, морфологические категории во множества аффиксов получим следующий список:

J={-нын,- га,- ны,-да, -дан } множество падежных аффиксов (Noun Cases);

T={-ым,-ың, -ыңыз, -сы, -ы, -быз, -ңар, -ңыздар} множество притяжательных аффиксов (Possessive);

K={-лар} множество аффиксов множественного числа (Pl);

Zh={ -мын, -быз, -сың, -сыңар, -сыз, -сыздар} множество аффиксов лица (Personal);

Ch={ -ды, -ган, -ыптыр, -чу,...} множество аффиксов времени (Verb Tenses);

In={ -са, -гай, ...} множество аффиксов наклонения (Imperatives);

Neg={ба} множество аффиксов отрицательного значения (аспект negative категории Verb Tenses);

Q={бы} множество аффиксов вопросительного значения (аспект interrogative *категории Verb Tenses).*

Если скажем, что *Um* — это множество словоизменительных аффиксов, то он состоит из следующих частей:

Um={J, T,K,Zh,Ch, In,Neg,Q}

Правила соединения аффиксов

Именительными словами называем имя существительное, имя числительное, имя прилагательное, местоимение.

Определение 2: Если $U \in (Z \vee C \vee San \vee At)$, то как показано в (4) формуле *U+Um*, *U+Um+Km* сумма не выполняется, иначе говоря после словоизменительных аффиксов словообразовательные аффиксы не соединяются.

Также сохраняются и правила словоизменительных аффиксов:

$$Um=K+T+J+Zh+Q \tag{6}$$

На основе формулы (6) *U∈(Z∨C∨San∨At) для времени (4) формулу можно написать так.*

$$S=U+Km+K+T+J+Zh+Q, \quad (7)$$

В этой формуле некоторые элементы множества словоизменительных аффиксов могут быть равны свободным аффиксам.

Например:

S=’аталар’, U=’ата’, Km=∅, Um=K=’лар’;

S=’аталарыбыз’, U=’ата’, Km=∅, Um=K+T=’лар’+’ыбыз’;

S=’аталарыбыздын’, U=’ата’, Km=∅, Um=K+T+J=’лар’+’ыбыз’+’дын’;

S=’аталарсыңар’, U=’ата’, Km=∅, Um=K+T+Zh=’лар’+’сыңар’;

S=’аталарыбызсыңар’, U=’ата’, Km=∅, Um=K+T+J+Zh=’лар’+’ыбыз’+’сыңар’;

S=’аталарыбызсыңарбы’, U=’ата’, Km=∅,
Um=K+T+J+Zh+Q=’лар’+’ыбыз’+’сыңар’+’бы’;

Все вышеназванное можно посмотреть в следующей Таблице.

Таблица.

1	*2*	*3*	*4*	*5*	*6*	*7*
Именные основы Например: ата, казан, жүрөк и др.	*-лар*	*-ым (-м)* *-ың (-ң)* *-ы (-сы)* *-ыбыз* *(-быз)* *-ыңар* *(-ңар)* *(-лары)* *-ныкы*		*-нын* *-га* *-ны* *-да* *-дан* *-сыз* *-сыздар* *-ыңыз* *-ыңыздар* *-(а)т* *(-ыш –а –т)*	*-мын* *-м* *-быз* *-сың* *-сыңар*	*-бы* *-чы*

Алгоритм морфологического анализа

Соответствующий алгоритм представлен на Рисунке.

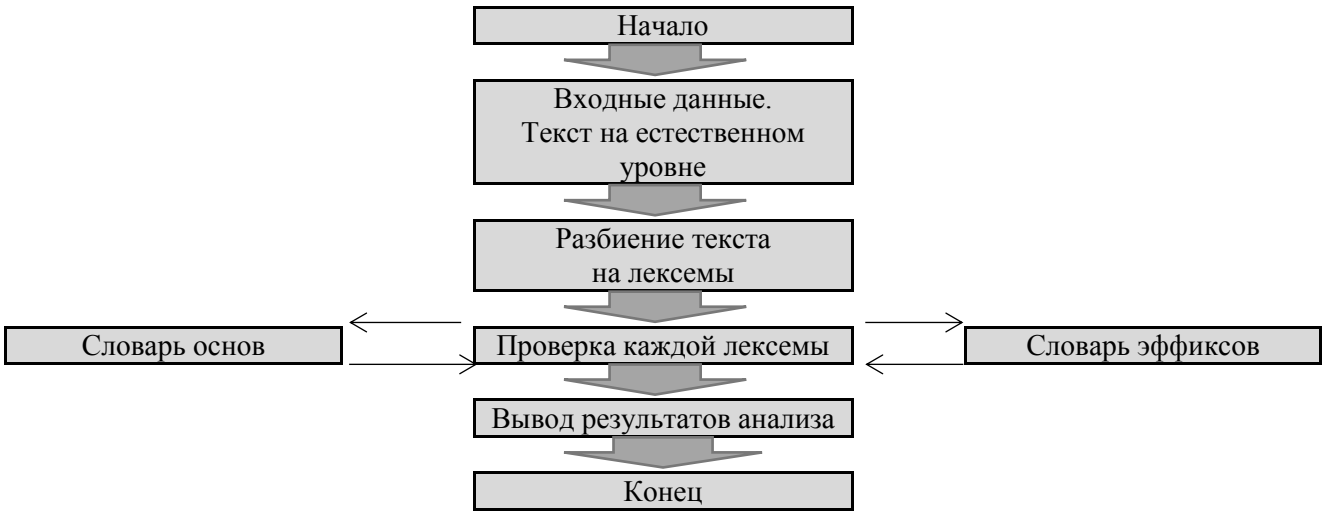

Рисунок. Алгоритм морфологического анализатора.

Заключение

Таким образом, первым шагом всех прикладных программ связанных с обработкой естественного языка является морфологический анализатор. Поэтому создание математической модели анализа является актуальной задачей. Вышеизложенной статье мы рассматривали модель стемминга и на основе этого алгоритм морфологического анализа текста. В дальнейшем мы будем использовать эти алгоритмы для создания машинного перевода.

Список литературы:

1. Садыков Т. Проблемы моделирования тюркской морфологии. Фрунзе: Илим, 1987. 103 с.

2. Панков П. С. Обучающая и контролирующая программа по словоизменению в кыргызском языке на ПЭВМ. Бишкек: Мектеп, 1992. 20 с.

3. Кочконбаева Б. О. О морфологическом анализе в приложениях автоматической обработки текста // Бюллетень науки и практики. 2018. Т. 4. №12. С. 608-612.

References:

1. Sadykov, T. (1987). Problemy modelirovaniya tyurkskoi morfologii. Frunze: Ilim, 103.

2. Pankov, P. S. (1992). Obuchayushchaya i kontroliruyushchaya programma po slovoizmeneniyu v kyrgyzskom yazyke na PEVM. Bishkek: Mektep, 20.

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

Работа поступила
в редакцию 11.02.2019 г.

Принята к публикации
16.02.2019 г.

Ссылка для цитирования:

Сатыбаев А. Д., Кочконбаева Б. О. Математическое моделирование и алгоритм морфологического анализа кыргызского языка // Бюллетень науки и практики. 2019. Т. 5. №3. С. 220-224. https://doi.org/10.33619/2414-2948/40/28.

Cite as (APA):

Satybaev, A., & Kochkonbaeva B. (2019). Mathematical modeling and algorithm of morphological analysis of the Kyrgyz language. *Bulletin of Science and Practice, 5*(3), 220-224. https://doi.org/10.33619/2414-2948/40/28. (in Russian).