

Комитет науки Министерства образования и науки
Республики Казахстан
РГП «Институт информационных и вычислительных технологий»
КН МОН РК



25 лет
Независимости
Республики Казахстан



25 лет
Институту
информационных и
вычислительных
технологий

МАТЕРИАЛЫ
Международной научной конференции
«Информатика и прикладная математика»
(«Computer science and Applied Mathematics»),
посвященной 25-летию Независимости Республики Казахстан и
25-летию Института информационных и
вычислительных технологий

Часть II

г. Алматы, 21-24 сентября 2016 года

Алматы
2016

© Институт информационных и вычислительных технологий, 2016

Секция 3. ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА		7
<i>Kylyshkanov M.K., Uvaliyeva I.M., Belginova S.A.</i>	Informational and analytical system of the biochemical analysis of blood in children	8
<i>Мамбетаев О.А., Мұсабаев Р.Р., Тасбалатулы Н.</i>	Қазақ тілінің дыбыстық түрленімінің ерекшеліктері	20
<i>Песитханова А.Е.</i>	Жеке тұлғаны дауысы бойынша тану жүйесінің жұмысына сипаттама	28
<i>Солтангельдинова М., Мансурова М.Е., Бердибеков С.</i>	Гиперграф негізінде кілттік сөздерді шығару	35
<i>Амиргалиев Е.Н., Мустафин С.А., Муратханова Т.А.</i>	Вопросы определения движения робота на плоскости	44
<i>Баракшин В.Б., Кожемякина О.Ю., Забайкин А.В., Хайтова В.Д.</i>	Автоматизация процесса создания метрических справочников и конкордансов с использованием компьютерных алгоритмов анализа поэтических текстов на русском и казахском языках	47
<i>Кетининов Ч., Амиргалиев Е., Арсланов М.</i>	Сравнение производительности слоев сверточных сетей для задачи визуального определения местоположения	55
<i>Койбағаров К.Ч., Мұсабаев Р.Р.</i>	Автоматическая классификация отзывов с использованием методов машинного обучения	62
<i>Мустафин С.А.</i>	Моделирование процесса развития закладочного материала	71
<i>Нуртазин А.Т., Хисамиев З.Г.</i>	Автоматическое извлечение следствий из семантической структуры текста	75
<i>Тукеев У.А., Турганбаева А.</i>	Лексикон-фрейм стемминг для казахского языка	84
<i>Шарипбай А.А.</i>	Искусственный интеллект в Республике Казахстан	89
Секция 4. ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ И ЗАЩИТА ДАННЫХ		102
<i>Kostopoulos George</i>	An original numerical factorization algorithm	103

ЛЕКСИКОН-ФРИ СТЕММИНГ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Тукеев У.А., Турганбаева А.

e-mail: ualsher.tukeyev@kaznu.kz

КазНУ им. аль-Фараби, ДГП НИИ ММ, Казахстан

Аннотация. В работе предлагается новый подход к построению алгоритма стемминга без словаря (лексикон-фри) для казахского языка на основе полного множества окончаний казахского языка.

1 Введение

Вопросы обработки текстов казахского языка слабо проработаны. Это связано во-первых, со спецификой казахского языка, как языка со сложной морфологией, и во-вторых, с еще недостаточно активным исследованием казахского языка в этой области. Однако, вопросы обработки текстов на казахском языке на практике весьма актуальны. Одним из актуальных вопросов является область информационного поиска данных в казахоязычных текстах. В данной области важным вопросом является проблема быстрого поиска заданных слов в документах. Одним из способов быстрого поиска слов является поиск по основам слов среди ключевых слов документов, что позволяет выбрать соответствующий документ как искомый. В связи с этим, одним из путей решения поиска по основам слов является выделение в слове основы, а именно, стемминг. Стемминг слов можно делать на основе использования словарей или без использования словарей, а именно, лексикон-фри. Данный лексикон-фри подход, начатый Портером[1], получил активное развитие и применение в области информационного поиска.

В данной работе предлагается подход к построению лексикон-фри стемминг алгоритма для казахского языка на основе полного множества окончаний казахского языка. Полнота окончаний языка обеспечивает достаточно высокий уровень выполнения стемминга слов.

2 Предшествующие работы

В области разработки алгоритмов стемминга представлено достаточно много работ. Хотелось отметить некоторые из них, достаточно близкие к данной работе, во-первых, по исследуемому языку, и во-вторых, по подходу к построению самого алгоритма. В связи с этим, это работа самого Портера, недостатком которого является достаточно высокий процент ошибок. В работе Сегаловича [2] предлагается эффективный алгоритм стемминга для русского языка с использованием как особенностей системы окончаний русского языка, так и использования словаря, что повышает качество стемминга. В работе Федотова и др. предлагается алгоритм лемматизации для казахского языка, в котором рассмотрена систематизация окончаний казахского языка не обладающая полнотой[3].

3 Полное множество окончаний казахского языка

В казахском языке окончания подразделяются на два больших класса:

- окончания на именной основе, к которым относятся окончания существительных, прилагательных и числительных;

- окончания на глагольной основе, к которым относятся окончания глаголов, причастий, деепричастий, наклонений и залогов.

В работе [4] впервые было рассмотрено построение множества окончаний казахского языка и возможность его использования для эффективного обучения казахскому языку.

Базовые аффиксы казахского языка, из которых формируются окончания казахского языка могут быть следующих четырех типов:

- аффиксы множественного числа (обозначены через K),
- аффиксы притяжательные (обозначены через T),
- аффиксы падежные (обозначены через C),
- аффиксы личные (обозначены через J).

В работе [5] рассмотрены всевозможные размещения варианты размещений типов аффиксов в окончаниях слов казахского языка: из одного типа, из двух типов, из трех типов и из четырех типов. Число размещений определяется формулой: $A_n^k = n!/(n-k)!$.

Тогда, количество размещений будет определяться следующим образом:

$A_4^1 = 4!/(4-1)! = 4$, $A_4^2 = 4!/(4-2)! = 12$, $A_4^3 = 4!/(4-3)! = 24$, $A_4^4 = 4!/(4-4)! = 24$. Всего возможных размещений 64.

В работе [5] путем рассмотрения семантически допустимых размещений определено полное множество семантически допустимых типов окончаний казахского языка.

Так, для слов с именными основами семантически допустимое количество типов окончаний казахского языка равно 15. Семантически допустимое количество типов окончаний казахского языка для причастий равно 11, для глаголов равно 25, для деепричастий равно 1, для наклонений равно 6 и для залогов – 8. Итого, общее количество типов окончаний слов с глагольными основами будет 51.

Итого, общее количество окончаний с именными основами плюс общее количество типов окончаний слов с глагольными основами будет равно 66.

По данным типам окончаний построены конечные множества окончаний для всех основных частей речи казахского языка. Так, для частей речи с именными основами количество окончаний равно 1213 (учтены все варианты множественного числа), а количество окончаний частей речи с глагольными основами составляет: глаголы – 432, причастия – 1582, деепричастия – 48, наклонения – 240, залогов – 80. Итого, 3565 всего окончаний.

4 Алгоритм стемминга без словаря для казахского языка

Принцип предлагаемого алгоритма стемминга слов казахского языка, основанного на полной системе окончаний заключается в следующем:

В системе окончаний казахского языка все окончания разбиваются на классы по длине окончаний. В слове сначала ищется окончание максимальной длины для данного слова: оно будет на два символа меньше длины слова (предполагается, что основа не может длины меньше, чем 2). Предполагаемое окончание длины L ищется в соответствующем классе окончаний длины L. Если окончание не находится в данном классе, то длина предполагаемого окончания уменьшается на единицу и ищется в соответствующем классе окончаний и т.д. до тех пор, пока не найдется окончание или слово будет без окончания.

Обозначения:

$L(e)^{max}$ - есть максимальная длина окончаний в системе окончаний языка;

w - анализируемое слово;

$e(w)$ - окончание анализируемого слова;

$L(w)$ - длина анализируемого слова $L(w)$;

$L[e(w)]$ - предполагаемая длина окончания данного слова.

$L[e(w)]^{max}$ - максимальная длина окончания данного слова.

Шаги алгоритма:

1. определяется длина анализируемого слова $L(w)$.

2. определяется максимальная длина окончания анализируемого слова:

$$L[e(w)]^{max} = L(w) - 2.$$

где 2 - есть минимальная длина основы слова.

3. Если $L(w) \leq L(e)^{max}$ (если длина слова w меньше или равно максимальной длины окончаний в системе окончаний языка), то предполагаемой длине окончания данного слова $L[e(w)]$ присваиваем значение максимальной длины окончания анализируемого слова:

$$L[e(w)] = L[e(w)]^{max}. \text{ Далее перейти на 5.}$$

4. Иначе: предполагаемой длине окончания данного слова $L[e(w)]$ присваиваем $L(e)^{max}$:

$$L[e(w)] = L(e)^{max}.$$

5. Сделать выборку окончания $e(w)$ длины $L[e(w)]$ из данного слова w .

6. Проверка $e(w)$ на совпадение с окончанием из списка окончаний длины $L[e(w)]$.

Если совпадает, то определяем основу данного слова:

$st(w) = w - e(w)$, т.е. из данного слова выделяется основа.

Иначе

7. Уменьшаем предполагаемую длину окончания данного слова на единицу:

$$L[e(w)] = L[e(w)] - 1.$$

8. Если $L[e(w)] < 1$, то слово w без окончания. Переход на п.8.

Иначе - переход на п.5

9. Конец

5 Экспериментальные результаты

Для того чтобы проверить работу алгоритма и программы было сделано следующее тестирование. Было выбрано несколько слов и к этим словам добавлялись все возможные окончания. Например: рассмотрим слова «адам», «нышпан» и «пешим», для которых построен массив производных слов, добавляя к ним все возможные окончания. С помощью программы находим основы из слов. Полученные результаты показывают, что точность поиска основы составляет примерно 85%. Для слова «адам» было сформировано 110 производных слов, для слова «пешим» 78 производных слов, а для слова «нышпан» 86 производных слов. В таблице 1 рассматриваются ошибки программы при неправильном определении основы.

Таблица 1 – Причины неправильного нахождения основы

Виды окончания	Поиск основы	Причина неправильности основы
Окончания из одной буквы	Возможно неправильное нахождение основы	Из-за выбора падежных окончаний вместо притяжательных окончаний
Окончания из 2 и 3 букв	Возможно неправильное нахождение основы	В связи с приятием последних букв основ (например, -м, -ым, -ім), как притяжательное окончание.
Окончания из 4, ..., 10 букв	Основа находится правильно	

6. Дальнейшие работы

Дальнейшие работы связаны с совершенствованием алгоритма и программы:

1) добавление функций преобразования для слов с изменяющейся основой в силу правил сингармонизма казахского языка;

На казахском языке имеются буквы, которые при добавлении окончания меняются на другие буквы. Например, если слова заканчиваются на буквы «к, қ, п», то при добавлении окончания соответственно они меняются на буквы «г, ғ, б».

Таблица 2 – Преобразования основ

Слова	Основы	Окончания
Бірілігі	Біріліг	г – к + і
бостандығы	Бостандығ	г – қ + ы
Мектебі	Мектеб	б – п + і
Шегіне	Шег	г – к + іне
бірілігінің	Біріліг	г – к + інің
аумағының	Аумағ	ғ – қ + ының
оңтүстігінде	Оңтүстіг	г – к + інде

3. поиск решение следующих проблем:

а) на казахском есть слова которые при добавлении окончания одна буква выпадает. Например, рассмотрим слово «халық» с выпадением буквы «ы».

Таблица 3 – Преобразования основ

Слово	Основа	Окончание
Халқына	Халқ	ына

б) поиск решения проблем принятия последних букв основ за притяжательное окончание. Например, рассмотрим следующую таблицу.

№	Основа + окончание	Результат программы поиска основ
1	қала + м (мой город)	қала
2	қалам + без окончания (ручка)	қала
3	қала + мды	қала
4	қалам + ды	қала

Здесь, в первом и в третьем слове основа найдено правильно, а во втором и в четвертом словах неправильно. Программа принимает слово «қалам» как «қала» + «м» (притяжательное окончание), и в основу выводит слово «қала».

Рассмотрение вышеуказанных вопросов будет способствовать улучшению качества предлагаемого алгоритма и программы лексикон-фри стемминга слов казахского языка.

Литература

1. Porter M. F. An algorithm for suffix stripping // Program. 1980.
2. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine Yandex // iseg@yandex-team.ru.
3. Федотов А.М., Тусупов Д. А., Самбетбаева, М. А., Еримбетова А. С., Бакиева А. М., Идрисова А. И. Модель определения нормальной формы слова для казахского языка // Вестн. Новосиб. Гос. ун-та Серия: Информационные технологии. 2015. Т. 13, вып. 1. С. 107-116.
4. Бектасов К., Ахабасов А., Керимбаев Е., Молдабеков К. Краткий казахско-русский словарь. Приложение 1- Список окончаний казахского языка. - Алма-Ата: Главная редакция Казахской советской энциклопедии, 1991, -256 стр.
5. Tukeyev, U., Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. Proceedings of the international conference "Turkic languages processing" TURKLANG-2015 September 17-19, Kazan, Tatarstan, Russia, 2015. 91-100 pp