Компьютерный анализ текстов

Методы лингвистического анализа текстов на татарском языке и их применение в поисковой системе Exactus*

А. Р. Гатиатуллин, А.М. Баширов, Г.С. Осипов, И.В. Смирнов, А.О. Шелманов

Аннотация. В работе описываются методы и технологии, использованные при разработке модулей морфологического анализа тюркских словоформ. Технологии в основе этих модулей универсальны для всех тюркских языков, однако на данном этапе проекта реализованы программные модули, производящие обработку только татарских словоформ. Эти модули встроены в информационно-поисковую систему Exactus, которая ранее поддерживала работу с текстами только на русском и английском языках. Внедрение новых программных модулей позволит увеличить количество языков, анализируемых поисковой системой Exactus, а использованные в них технологии представляют интерес для работы с языками агглютинативного типа, в число которых входят все языки тюркского семейства.

Ключевые слова: *информационно-поисковая система, морфологический анализатор, татарский язык.*

Введение

В настоящее время существует большое количество систем, выполняющих анализ и обработку текстов на нескольких языках, в их число входят многоязычные поисковые системы. Одной из таких поисковых систем с возможностью многоязычного поиска является система Exactus**. Ранее эта система позволяла производить поиск только на русском и английском языках. В рамках нашего исследования была поставлена задача добавить функции поиска на языках тюркской группы.

Языки, входящие в тюркскую группу, структурно достаточно близки между собой, поэтому реализация в поисковой системе возможности работать с одним из тюркских языков позволит разработать технологии, которые без существенных изменений могут быть использованы для других языков тюркской группы. Тюркским языкам присущи следующие свойства, отличающие их от индоевропейских языков: автоматная левосторонняя морфология, агглютинация, отсутствие жесткой границы между парадигматическими классами, потенциально неограниченный объем парадигмы, нежесткое распределение лексики по грамматическим классам и частям речи.

Основными программными модулями для поисковой системы Exactus, разрабатываемыми в рамках нашего исследования, являются модуль графематического анализа и модули морфологического анализа с использованием и без использования словаря основ. В этой работе описывается создание модулей морфологического анализа для татарского языка и их вне-

* Работа выполнена при поддержке РФФИ, проект №13-07-00494 «А»

** http://exactus.ru/

дрение в поисковую систему Exactus. Работа выполнялась в рамках проекта по разработке комплексных моделей данных на основе ситуационного анализа текстов в задачах многоязычного поиска.

1. Подходы к морфологическому анализу тюркских языков

В настоящее время существует два основных подхода к созданию модулей морфологического анализа:

- парадигматический [1, 2];
- автоматный [3, 4].

В парадигматическом подходе используются два словаря: словарь основ и словарь парадигм. Принцип работы морфологических анализаторов, базирующихся на этом подходе, состоит в том, что каждой лемме в словаре присваивается индекс типа парадигмы, который отсылает к списку образцов парадигм (рис. 1). Парадигматический подход чаще всего используется для анализа флективных языков, в число которых входит и русский язык. Во флективных языках размеры парадигм невелики, но зато велико количество этих парадигм. В базе данных анализатора хранится полная парадигма для каждого типа основ. Этот подход можно встретить также и в морфологических анализаторах для тюркских языков в коммерческих системах, например, в продуктах ABBYY или Microsoft (в частности, встроенный грамматический корректор Microsoft Office). Парадигматический подход использован в этих системах для тюркских языков с той целью, чтобы не вносить изменения в их программное ядро.

Учитывая струкурные особенности тюркских языков, такие как автоматная морфология и неограниченность парадигмы, парадигматический подход для работы с ними не всегда эффективен. В работе [5] также отмечается, что для тюркских языков характерен еще целый набор отличительных свойств:

- развитая система грамматически однозначных аффиксальных морфем, где, как правило, один аффикс выражает один грамматический признак (хотя есть и исключения, например, аффиксы притяжательности);
- в отличие от флективных языков в тюркских языках отсутствуют различные парадигматические классы в рамках одного словоизменительного типа;
- отсутствие значимых чередований в основах, четкая фонетическая обусловленность использования алломорфов.

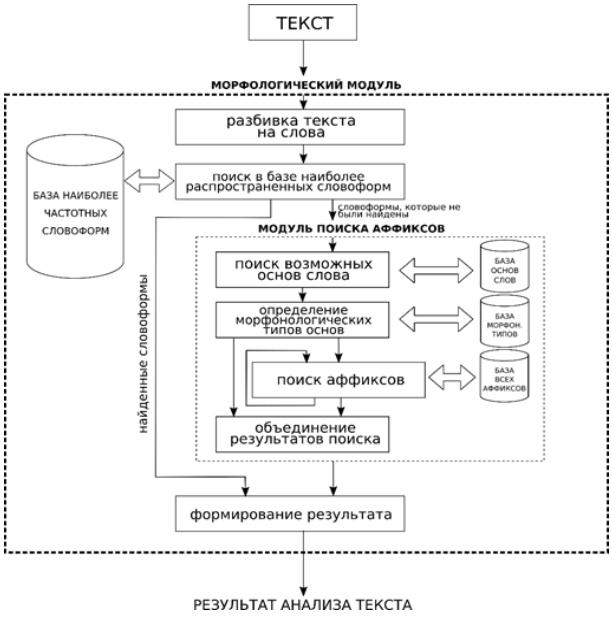

Рис. 1. Алгоритм морфологического анализатора

В настоящее время морфологические анализаторы реализованы для многих тюркских языков: татарского [6], башкирского [7,8], казахского [9], чувашского [10], турецкого [4], хакасского [5] и др.

Авторы [5] выделяют в хакасском языке три основных грамматических класса: имена, глаголы и неизменяемые частицы, послелоги, союзы и т.п. В морфологическом анализаторе, предлагаемом в нашем проекте, выделяются следующие морфологические типы: имя (N), глагол (V), признак (A), неизменяемая часть речи (S). Также нами выделен отдельный подтип имен – имена собственные (PN), который задается двумя словарями: имен и географических названий. Такое разделение на морфологические типы обусловлено только правилами морфотактики и не затрагивает синтаксических и семантических особенностей этих морфологических типов.

В работе [5] отмечается, что дифференциация между грамматическими классами выражена слабо, особенно между разрядами имени: слово может трактоваться как существительное, прилагательное или наречие в зависимости от его синтаксической функции:

- существительное может выступать в роли определения: «***таш*** *йортлар»* (рус. «каменные дома»);
- прилагательное может выполнять в предложении любую функцию: «олыга *урын бир»* (*рус.* «старшему место уступи»).

2. Модули морфологического анализа текстов на тюркских языках

2.1. Модуль морфологического анализа, использующий словарь основ

На более ранних этапах проекта нами уже был разработан вариант морфологического анализатора тюркских словоформ. Однако он не устраивал по скорости обработки текста, поэтому потребовалась доработка этого программного модуля. Особенностью указанного анализатора является то, что программная часть является универсальной для всех тюркских языков, а вся информация, необходимая для анализа словоформ конкретного языка, находится в базе данных, которая состоит из двух словарей: словарь основ и словарь морфотактических правил следования алломорфов в словоформе.

В качестве инструментария для разработки системы морфологического анализа словоформ была выбрана система управления базами данных MySQL.

В базе данных реализованы следующие группы таблиц:

1) Морфотактические правила:
   - таблица частей речи;
   - таблица значений морфем;
   - таблица склонений слов;
   - таблица аффиксов-алломорфов;
   - таблица связей аффиксов между собой (связь многие-ко-многим);
   - таблица связей склонений слов с аффиксами (связь многие-ко-многим).
2) Словарные данные:
   - таблица со списком слов языка (словарь основ).

В базе данных для каждого элемента из словаря основ указан его морфонологический тип. Этот тип определяет набор алломорфов, которые могут следовать сразу за любой основой данного типа, а также какие алломорфы могут следовать далее в словоформе. При выполнении анализа словоформы анализатор определяет возможные сочетания основ с последующими аффиксами, а затем возвращает полученные результаты в виде массива со следующими параметрами: основа с указанием ее уникального идентификатора (идентификатор необходим для связей с другими словарями), морфологический тип, набор аффиксов и морфологических категорий, которые соответствуют этим аффиксам.

Из-за того, что основа в зависимости от прибавляемого аффикса может частично меняться, в таблице со списком слов хранятся максимальные неизменяемые части основ, а изменяемые части отнесены в таблицы с морфонологическими правилами. Поскольку было принято решение не указывать в схеме нулевых аффиксов, возникает ситуация, когда слово, стоящее, скажем, в начальной форме, не имеет присоединяемого аффикса, который должен вернуть также недостающую часть основы. По этой причине в словарь основ также были добавлены полные формы слов, которые не имеют аффиксов. Для них указывается связь с соответствующей неполной основой.

Схема работы алгоритма морфологического анализатора изображена на рис. 1.

Рассмотрим процесс обработки слова морфологическим анализатором. Сначала производится поиск возможных основ. Переданное слово разбивается на части, которые гипотетически могут соответствовать основам в словаре, например, для словоформы «кешеләргә» (рус. «людям») будут образованы следующие возможные варианты основ: «к», «ке», «кеш», «кеше», «кешел» и т.д. Стоит отметить, что в конфигурации системы можно задавать максимально возможную длину основы, чтобы уменьшить число сравниваемых вариантов. После этого производится поиск основ в словаре для полученных вариантов. Исходя из найденных основ определяется, какие склонения им соответствуют, и затем производится поиск аффиксов для оставшейся части слова по аналогичному принципу, но уже в таблице аффиксов-алломорфов. Сначала ищется аффикс склонения, например, для основы «кеше» это будет поиск совпадений с «л», «лә», «ләр», «ләрг» и «ләргә», затем для каждого найденного аффикса на основе таблицы связей аффиксов производится поиск последующих возможных аффиксов и так далее до тех пор, пока не окончится остаточная часть слова (если она вообще была). Исходя из полученных аффиксов и основ формируется результирующий массив с указанием для каждого элемента идентификатора слова, его части речи и морфологических значений найденных аффиксов (если они есть).

Морфологический анализатор выдает следующую информацию:

- цепочку алломорфов;
- цепочку аффиксальных морфем;
- цепочку морфологических категорий;
- комбинированный вариант в виде набора аффиксов и морфологических категорий.

Результат работы анализатора для нашего примера будет выглядеть следующим образом:

[Основа: «кеше» (N), окончание: PL (-ЛАр) + DIR (-ГА)].

Здесь окончание содержит как обозначения морфологических категорий, так и обозначения

аффиксов, которые используются для обозначения этих морфологических категорий.

В данном примере PL – обозначение морфологической категории множественности, в татарском языке эта категория выражается с помощью аффикса ЛАр.

DIR – обозначение морфологической категории, называемой директивом, в татарской грамматике эта морфологическая категория называется направительным падежом. Для выражения этой морфологической категории в тексте в татарском языке используется аффикс ГА.

Такая система обозначений морфологических категорий отличается от системы обозначений морфологических категорий русского языка, поскольку для тюркских языков свойственно то, что для обозначения одной морфологической категории используется, как правило, одна аффиксальная морфема.

Для улучшения производительности системы были использованы следующие методы:

- сеансовое кэширование результатов некоторых запросов в программе поиска;
- перенос используемых таблиц с диска в оперативную память: в случае с MySQL – это изменение типа (engine) таблицы на «MEMORY»;
- в качестве индекса для таблиц в оперативной памяти использовался преимущественно тип индекса HASH, который позволяет выполнять сравнение для операций равенства/неравенства.

Для дальнейшей оптимизации был составлен тестовый текст из случайных слов из словаря, к которым не требуется добавление аффиксов. На этом тексте морфологический анализатор был протестирован в двух режимах: с включенным поиском аффиксов и без поиска аффиксов. Тестирование показало, что при отключенном поиске аффиксов программа работает в 10-18 раз быстрее.

На основе полученных результатов было принято решение закэшировать в базе данных наиболее распространенные формы слов для каждой основы. Под формой слов подразумевается сочетание одного или нескольких аффиксов. В систему были внесены следующие изменения: сначала производится поиск словоформы в базе часто встречающихся форм слов нашего словаря, и только в случае, если форма не найдена, производится анализ этого слова. Таким образом, в базу данных были добавлены таблицы часто встречающихся словоформ (количество записей в 100 раз больше таблицы словаря) и морфологических значений соответствующих словоформ.

Наиболее распространенные формы слов были выбраны на основе автоматического анализа небольшого корпуса текстов (около 6-8 тыс. текстов различных стилей). Поскольку ряд часто встречающихся слов отсутствовал в нашем словаре, мы приняли решение добавить также и их в таблицу распространенных словоформ с указанием, что определение словоформы данного слова не является возможным. В результате скорость работы программы выросла в 10-12 раз. Однако, если использовать только распознавание часто встречающихся словоформ, то скорость увеличивается до 18 раз, но при этом количество проанализированных слов снижается на 2-5%.

2.2. Морфологический анализатор без словаря основ

В текстах на тюркских языках встречается много слов, которых нет в базовых лексических словарях. Это различные имена собственные или заимствования из других языков. Строить для подобных словоформ отдельные словари нецелесообразно, поскольку такие слова могут встретиться лишь однажды в одном конкретном тексте. С целью обеспечения возможности семантического поиска для предложений с такими словоформами также бывает необходимо построить описание ситуации. Поскольку для выражения отношений между объектами ситуации используются аффиксальные морфемы, то для выделения этих отношений требуется определять категориальную принадлежность основы словоформы, а также набор его аффиксов.

В результате работы морфологического анализатора без использования словаря основ, количество варинатов анализа будет намного больше, чем при работе со словарем. При выдаче результатов анализа программа автоматически производит ранжирование выдаваемых вариантов анализа. Результатами с наибольшим рангом будут цепочки аффиксальных морфем с наибольшим числом аффиксов.

Например, получая на вход словоформу «Иванныкыларга», программа выдаст следующие варианты анализа:

N (*Иван*) – нЫкЫ – Лар – ГА	ранг 3
N (*Иванныкы*) – Лар – ГА	ранг 2
N (*Иванныкылар*) – ГА	ранг 1
V(*Иванныкыла*) – ЫРГА	ранг 1

Для реализации анализатора использована уже существующая база данных. Тем не менее, в рассматриваемом случае пропадает необходимость в таблице основ.

В модуле морфологического анализа без поиска основ реализован алгоритм обратного поис-

ка. Рассмотрим тот же пример: словоформа «кешеләргә». Сначала производится создание списка возможных вариантов аффиксов с конца слова: «ә», «гә», «ргә», «әргә» и т.д. Количество вариантов ограничено максимально возможной длиной аффикса, задающейся в конфигурации. Поиск возможных вариантов аффиксов ведется в таблице аффиксов-алломорфов. Затем для каждого найденного аффикса проводится проверка оставшейся части слова с учетом дальнейших возможных аффиксов (на основе таблицы взаимосвязей аффиксов). В результате обязательно должна остаться какая-то часть слова, которая не была распознана (поскольку не бывает слов без основ). После того, как остается основа, исходя из найденных сочетаний аффиксов, мы можем определить, какая часть речи и какое склонение может соответствовать каждому возможному случаю распознавания. Исходя из возможных склонений слов, мы можем восстановить возможные их основы. Например, для словоформы «китабыма» (рус. «для моей книги») в качестве основы система предложит как вариант «китаб», так и вариант «китап». Второй вариант является правильным, однако система не может этого знать, поскольку не использует словарь основ.

В качестве дополнительного элемента уточнения результата мы также выполняем сравнение окончания основы с возможными окончаниями основ данного склонения. Например, для словоформы «этапта», основой которой является слово «этап», проверяется, что склонение, соответствующее аффиксу «-та», применимо лишь к словам, завершающимся на глухой согласный звук, в противном случае аффикс должен быть «да». Таким образом, для записи «этапда» основа «этап» будет невозможной.

Для повышения эффективности работы данного модуля может потребоваться дальнейшее уточнение возможных окончаний основ склонений, а также кэширование в оперативной памяти наиболее распространенных сочетаний аффиксов с применением оптимизированных для данной системы индексов.

3. Поиск на татарском языке в системе Exactus

Лингвистический процессор для языков тюркской группы в составе модулей графематического и морфологического анализа был внедрен в поисково-аналитическую систему Exactus [11]. Результатом внедрения стала разработка поисковой машины по татарским сайтам.

В основе системы Exactus лежит метод реляционно-ситуационного поиска [12, 13, 14]. Его особенность заключается в том, что при поиске сравниваются не только ключевые слова запроса и документа, но также синтаксические деревья и неоднородные семантические сети, отражающие семантическую структуру предложений: значения синтаксем и семантические отношения на значениях синтаксем. Для построения сети необходимо выполнить глубокий лингвистический анализ текста, включая синтаксический и семантический. Однако лингвистический процессор для языков тюркской группы на данный момент не позволяет проводить эти виды анализа, поэтому он не позволяет реализовать все возможности системы Exactus. Тем не менее, процессор может проводить лемматизацию, определять морфологические характеристики слов, а также определять лексические значения слов. Лексическое значение в данном случае выражается группой синонимичных слов, которые образуют синсеты в словаре. Эти группы объединяют схожие по смыслу слова как в одном языке, так и в разных тюркских языках: татарском, башкирском, хакасском и др. Таким образом, возможности лингвистического процессора позволяют реализовать поиск по леммам ключевых слов (и их характеристикам), а также поиск по лексическим значениям слов (поиск с учетом синонимов). Например, в текущей реализации при запросе «акылыңны» поисковая система найдет не только документы, содержащие леммы «акыл», но и релевантные документы, в которых встречается «интеллект», «аң» и др. Поскольку в синсетах содержатся слова на разных языках тюркской группы, это также открывает возможность кросс-языкового поиска по лексическим значениям слов.

Тестовая версия поисковой машины работает на данный момент не со всеми языками тюркской группы, а только с татарским. Важной особенностью татарских веб-ресурсов является то, что многие страницы на этих ресурсах содержат информацию также и на русском языке. Кроме того, документы этих ресурсов часто содержат отдельные русскоязычные слова, отсутствующие в татарском языке. Поэтому в поисковой машине была проведена интеграция лингвистических процессоров для татарского и русского языков. В частности, был реализован компонент, который определяет язык слова и, таким образом, позволяет выбрать анализатор для его обработки. Если по форме слова и его контексту слово нельзя однозначно отнести к татарскому или к русскому языку, то создается два омонима, которые затем учитываются при поиске.

Рассмотрим процесс индексации докумен-

тов в разработанной поисковой машине. На первом этапе документы загружаются из Интернет при помощи сетевого краулера. Краулер рекурсивно обходит ссылки на страницах, фильтруя повторяющиеся и нерелевантные ресурсы, а также извлекает метаданные из страниц по заранее созданным шаблонам. Краулер работает с большинством современных Интернет-протоколов (http, ftp, https и др.) и позволяет обходить скрытый веб. Далее из загруженных документов извлекается непосредственно текст. Система может обрабатывать основные текстовые форматы: html, pdf, doc, docx, rtf, ps. Кроме того, при необходимости автоматически включаются модули распознавания текста по изображениям, что в основном требуется для обработки документов в формате pdf или ps с некорректной или отсутствующей текстовой подложкой. Разметка исходного документа (html, xml) также запоминается и используется для определения значимости текстовых фрагментов. Извлеченный текст направляется модулям лингвистического анализа. В поисковой машине по татарским сайтам выполняется определение языка отдельных слов, а также морфологический анализ с определением значений отдельных слов. В итоге извлеченная лингвистическая информация, теговая разметка документа, а также метаданные документа индексируются в поисковой базе, в основе которой лежат реляционно-ситуационные структуры данных [15]. Заметим также, что при индексации и поиске фильтруются стоп-слова, которые определяются как по словарям, так и по частям речи, определенным в результате морфологического анализа. Схема процесса индексации документа в поисковой машине представлена на рис. 2.

Для тестирования поисковой машины и проведения экспериментальных исследований была собрана коллекция текстовых документов на татарском языке. Основная часть документов была получена при помощи автоматического обхода трех ресурсов: татарской версии Википедии*, Татарской электронной библиотеки**, а также официального портала Республики Татарстан на татарском языке***. Суммарный объем индексной базы на данный момент составляет более 110 тысяч текстовых документов.

Тестовая версия поисковой машины доступна в Интернете по адресу http://tat-exactus.isa.ru/. Веб-интерфейс машины позволяет проводить поиск по указанным ресурсам на русском и татарском языках. Для найденных документов строятся аннотации, а также доступна ссылка на исходную страницу ресурса (рис. 3).

* http://tt.wikipedia.com
** http://kitap.net.ru/
*** http://tatarstan.ru/tat/

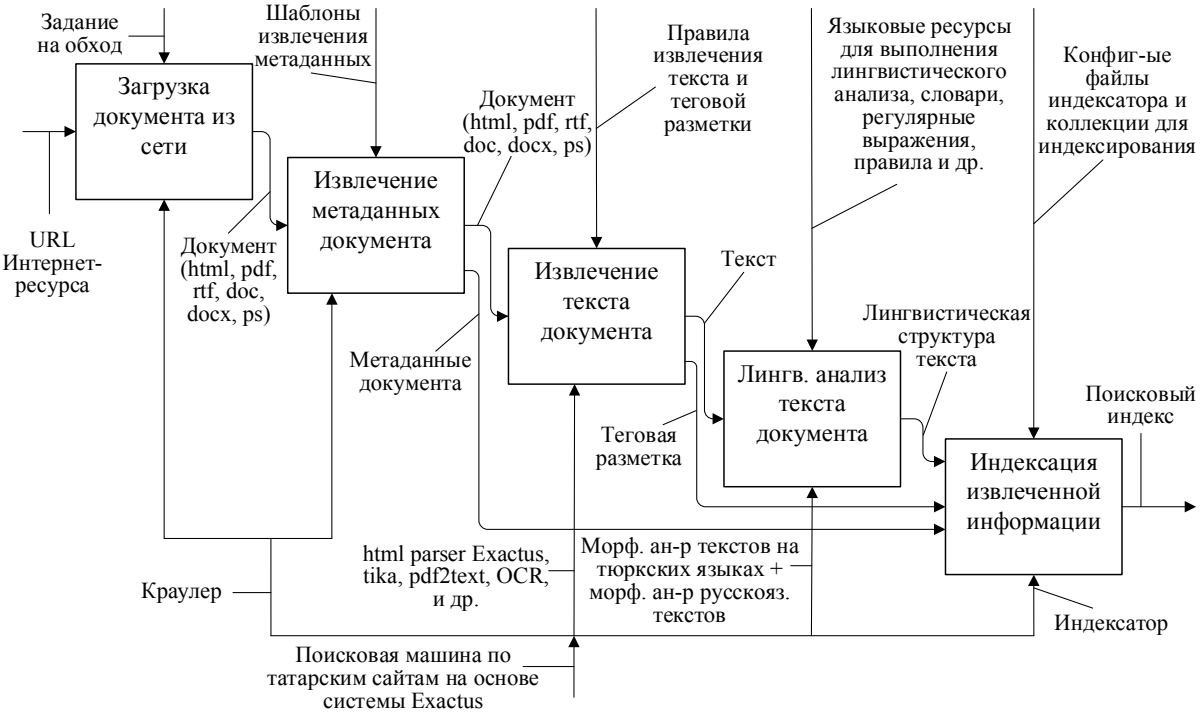

Рис. 2. IDEF0 – диаграмма процесса индексации документа в поисковой машине по татарским сайтам

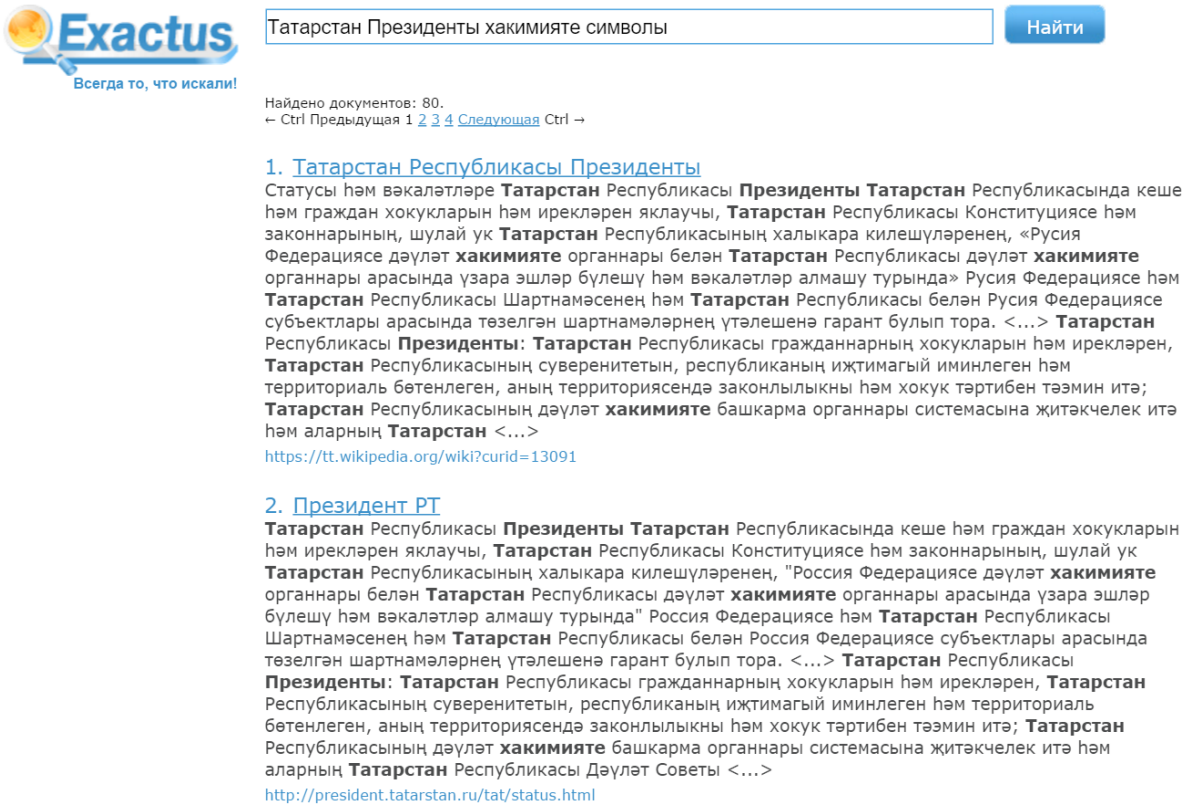

Рис. 3. Пользовательский веб-интерфейс поисковой машины по татарским сайтам

Заключение

Разработанные модули морфологического анализа татарских словоформ позволяют решать средствами технологий Exactus многие задачи обработки массивов текстов на татарском языке, включая поиск, классификацию, кластеризацию, реферирование текстов, поиск текстовых заимствований, извлечение информации из текстов и т.д. Представляет интерес возможность решения перечисленных поисково-аналитических задач на основе разработанных модулей и для других тюркских языков.

Дальнейшие работы будут вестись по нескольким направлениям. Во-первых, поскольку созданные модули морфологического анализа и архитектура системы позволяют работать со множеством языков, первоочередной задачей является реализация поддержки кросс-языкового поиска по документам на разных языках тюркской группы. Во-вторых, планируется расширение функций лингвистического процессора, добавление возможностей синтаксического и семантического анализа текстов на тюркских языках, что в дальнейшем позволит реализовать фразовый и семантический поиск. В-третьих, планируется расширение коллекций документов, по которым система сможет вести поиск. В них будут включены наиболее посещаемые Интернет-порталы (например, СМИ, социальные сети), а также крупные электронные каталоги.

Литература

1. *Тузов В. А.* Морфологический анализатор русского языка //Вестник СПбГУ, сер. 1. 1996. Вып. 1 (N15). С. 41–45.
2. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'99». Т. 2. С. 547–552. Казань, 1998.
3. *Antworth, E. L.* PC-KIMMO: a two-level processor for morphological analysis. Occasional Publications in Academic Computing No. 16. Dallas: Summer Institute of Linguistics, 1990, 273 p.
4. *Kemal Oflazer*. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, – Vol. 9, No 2, – 1994.
5. *Дыбо А. В., Шеймович А.В.* Автоматический морфологический анализ для корпусов

тюркских языков // Филология и культура – 2014. – №2.
6. *Сулейманов Д. Ш., Гильмуллин А. А., Гильмуллин Р. А.* База морфотактических правил для татарского глагола как основа двухуровневого морфологического анализатора // Сборник трудов Международного семинара «Диалог», Казань, 1998. – С. 597-609.
7. *Сиразитдинов З. А.* Алгоритмическая грамматика словоизменения башкирского языка // [Электронный ресурс]. URL: http://mfbl.ru/bashdb/algram/algram.htm (дата обращения: 19.09.2015).
8. *Орехов Б. В., Слободян Е. А.* Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 167–171.
9. *Шарипбаев А. А., Бекманова Г. Т., Ергеш Б. Ж., Бурибаева А. К., Карабалаева М. Х.* Интеллектуальный морфологический анализатор, основанный на семантических сетях // Материалы международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS-2012). Минск, БГУИР, 16–18 февраля 2012г. С. 397–400
10. *Желтов П. В.* Морфологический анализатор чувашского языка. Материалы Международной конференции студентов и аспирантов по фундаментальным наукам «Ломоносов 2002», М., 2002.
11. *Осипов Г. С., Тихомиров И. А., Смирнов И. В.* Семантический поиск в сети Интернет средствами поисковой машины Exactus // Труды одиннадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. Т3 - М.: ЛЕНАНД, 2008. - С. 323-328..
12. *Соченков И. В., Суворов Р. Е.* Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 1) // Информационные технологии и вычислительные системы. – 2013. – №2. – С. 69-78.
13. *Осипов Г. С., Смирнов И. В., Тихомиров И. А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. – 2008. – № 2. – С. 3–10.
14. *Relational–situational* method for intelligent search and analysis of scientific publications / Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov // Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13). – Vol. 968. – CEUR Workshop Proceedings, 2013.
15. *Соченков И. В., Суворов Р. Е.* Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 2) // Информационные технологии и вычислительные системы. – 2013. – №3. – С. 71-87.

Гатиатуллин Айрат Рафизович. Начальник отдела АН республики Татарстан. К.т.н. Окончил в 1994 г. Казанский ГУ. Количество печатных работ: более 40 (в т.ч. 1 монография). Область научных интересов: компьютерная лингвистика, тюркские языки. E-mail: agat1972@mail.ru

Баширов Артур Маратович. Зам. директора ООО «ТемирТех». Окончил в 2009 г. Казанский ГУ. Количество печатных работ: 5. Область научных интересов: компьютерная лингвистика.
E-mail: a.basheerov@gmail.com

Осипов Геннадий Семенович. Зам. директора ИСА ФИЦ ИУ РАН. Д.ф.-м.н., профессор. Количество печатных работ: более 180. Область научных интересов: представление знаний, приобретение знаний интеллектуальными системами, динамические интеллектуальные системы, семантический поиск. E-mail: gos@isa.ru

Смирнов Иван Валентинович. Зав. лабораторией ИСА ФИЦ ИУ РАН. К.ф.-м.н., доцент. Окончил в 2003 г. РУДН. Количество печатных работ: 50. Область научных интересов: компьютерная лингвистика, интеллектуальный анализ текстов, машинное обучение, интеллектуальный анализ данных, информационный поиск, информационные технологии в медицине. E-mail: ivs@isa.ru

Шелманов Артем Олегович. М.н.с. ИСА ФИЦ ИУ РАН. Окончил в 2011 г. НИЯУ МИФИ. Количество печатных работ:17. Область научных интересов: искусственный интеллект, компьютерная лингвистика, информационно-аналитические системы, машинное обучение. E-mail: shelmanov@isa.ru