СИСТЕМА СЛОЖНОГО МОРФОЛОГИЧЕСКОГО ПОИСКА В ПИСЬМЕННОМ КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

М. Р. Сайхунов, Р. Р. Хусаинов, Т. И. Ибрагимов
Казань

Аннотация. Работа посвящена описанию современного состояния Письменного корпуса татарского языка. В настоящей версии корпус способен выполнять морфологический анализ словоформы, производить фонетическую транскрипцию слов и предложений, а также представлять результаты анализа в письменной и устной формах языка.

Ключевые слова: корпусная лингвистика, татарский язык, обработка естественного языка, быстрый поиск, морфология, интернет, программирование.

Контактная информация: tatcorpus@gmail.com.

1. Письменный корпус татарского языка

Письменный корпус татарского языка функционирует в сети Интернет с 2012 года [1] и включает в себя самые разные тексты на татарском языке общим объемом более 116 миллионов словоупотреблений. Количество предложений в Корпусе превышает 10 миллионов.

С момента создания Письменный корпус татарского языка активно развивается. На сегодняшний день Корпус обладает и встроенной системой синтеза татарской речи, которая позволяет прослушивать как найденные в результате поиска предложения, так и произвольный текст любого объема. Авторы размещают различные дополнительные статистические материалы по мере их получения в результате обработки Корпуса (как при проведении собственных исследований, так и на основе поступивших внешних предложений), а именно:

списки наиболее часто употребляемых

- словоформ татарского языка;
- словосочетаний, состоящих из 2, 3, 4, 5 и 6 элементов;

списки частотностей

- лемм татарского языка, сгруппированные по частям речи;
- грамматических форм татарского языка;
- букв и их сочетаний в различных позициях слов;
- фонем и их сочетаний в пределах слова и ритмической группы.

2. Создание системы сложного морфологического поиска

В целях расширения функциональных возможностей Письменного корпуса татарского языка в 2014 году была произведена морфологическая разметка. Для этого использовалась разработанная международным проектом Apertium [4] система автоматической грамматической аннотации, которая поддерживает большое количество языков (в том числе и татарский).

Основными факторами в выборе системы Apertium являются:

- высокое качество морфологической разметки;
- наличие универсальной системы тегов для тюркских языков, что подразумевает перспективу выстраивания различных лексических и грамматических связей между корпусами разных тюркских языков;
- полная открытость исходных кодов и всех наработок (словари, правила).

В связи с появлением в корпусе большого объема новой метаинформации, в том же 2014 году начата работа над новой корпусной поисковой системой, удовлетворяющей следующим критериям:

- поддержка таких параметров поиска, как словоформа, лемма, грамматические (морфологические) теги, маска (шаблон), учет регистра (прописные и строчные буквы), расстояние между словами;

- возможность создания различных комбинаций на базе указанных параметров;
- легкий (простой) синтаксис запросов;
- высокая скорость поиска.

В ходе работ был учтен опыт различных известных проектов, среди которых TSCorpus [3], Sketch Engine [6] и др. [5]. Данная система в целом получила название «Сложный морфологический поиск», а в качестве ядра нами был разработан корпусный поисковый движок «**fastmorph**», который успешно справляется со всеми указанными выше задачами.

На сегодняшний день система сложного морфологического поиска активно используется в составе Письменного корпуса татарского языка, ведутся работы по внедрению дополнительных возможностей.

3. Возможности системы сложного морфологического поиска

Рассмотрим несколько примеров возможных запросов в системе сложного морфологического поиска. Следует помнить, что фигурные скобки здесь даны только для указания на соответствующие текстовые поля на странице поиска и не используются в реальных запросах!

Для начала произведем поиск по комбинации «**{<adj>} 1-2 {<n><dat>} 1-3 {<v><past>}**».

Рис. 1. Образец поискового запроса

Данное выражение означает, что первое слово должно быть *прилагательным* (<adj>), следующее за ним на расстоянии от одного до двух слов должно быть *существительным* (<n>) *в дательном падеже* (<dat>), и после него на расстоянии до трех слов должен идти *глагол* (<v>) в форме «*-ган* / *-гән* / *-кан* / *-кән*» *прошедшего времени* (<past>).

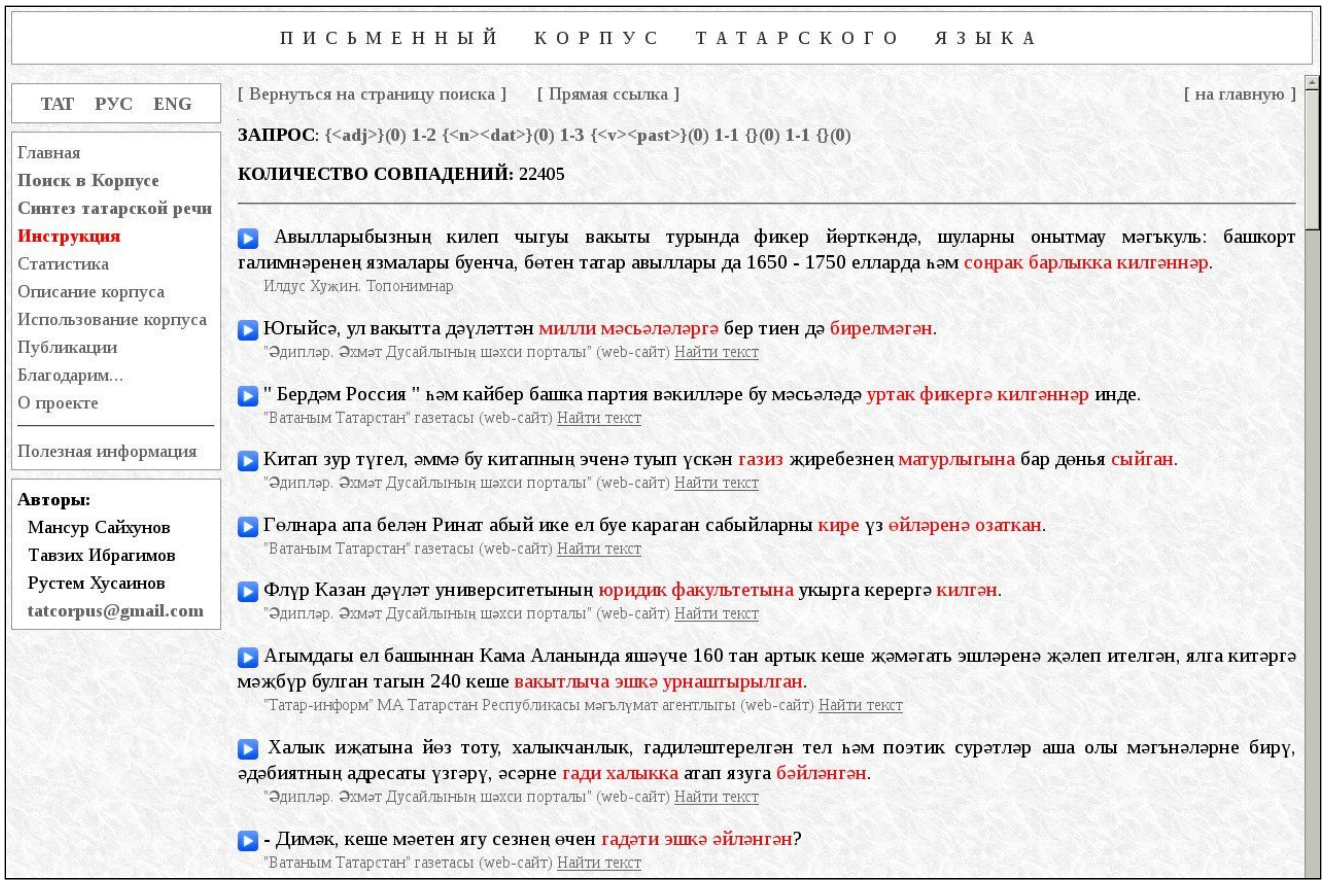

Рис. 2. Результат поискового запроса

В качестве второго примера укажем параметры «**{ил*} 1-1 {белән}**», которые означают, что первое слово должно начинаться на «*ил*», а следующее непосредственно за ним слово должно быть «*белән*» (*послелог в значении предлога* ‘с’).

Можно указать среднюю часть слова, используя запрос вида «**әме**», который соответствует словам «*керәмен, әмер, үсәме...*».

Шаблон поиска по концу слова выглядит как «**рны*». В результате получим предложения со словами «*дусларны, атларны, барны, кулларны...*».

Для поиска по началу, средней части и концу слова можно оформить запрос в виде «*к*әме*н*», что в итоге приведет к нахождению «*керәмен, каләмен, күләменнән, кияүдәмен...*».

Знак звездочка «*» совпадает с любым количеством (от нуля до бесконечности) любых символов, а знак вопроса («?») соответствует любому одиночному символу. Например, по образцу «*т?з**» будут найдены такие слова, как «*тиз, тозны, түзде, тазарды...*», но не «*тигез, тугызны, тәрәзә...*».

Все перечисленные поисковые параметры (словоформа, лемма, грамматические теги, шаблоны) могут быть комбинированы различными способами. Например, запрос «**{<prn>} 1-1 {(кеше)} 1-3 {ал*}**» будет искать все совпадения, где первое слово является *местоимением* (<prn>), следующее непосредственно за ним слово является одной из форм леммы «*кеше*» (‘человек’), и расположенное на расстоянии до трех слов слово, начинающееся на «*ал*».

В качестве еще одного примера рассмотрим следующую ситуацию. Допустим, что необходимо найти случаи употребления словоформы «***алма***» (*глагол,* означающий ‘не бери’), однако в результаты поиска попадут также предложения с омонимом «*алма*» (*существительное,* ‘яблоко’). Для того, чтобы исключить последние совпадения, можно поставить морфологический тег «<v>», определяющий данное слово как *глагол*: «***алма<v>***».

В этом случае есть другой способ решения данной проблемы. Можно вместе со словоформой «*алма*» указать соответствующую ей лемму, т.е. «*(ал)*» (‘брать’). В итоге запрос примет вид

«***алма(ал)***». Таким образом, система будет искать только те случаи «*алма*», где леммой данной словоформы является «*ал*» (‘брать’), опуская при этом все результаты, где лемма — «*алма*» (‘яблоко’).

Технически пользователь может даже оформить запрос в виде «*алма(ал)<v>*» или «*(ал)алма<v>*»… Система выполнит разбор данного выражения и будет искать примеры со словоформой «*алма*», которая представлена леммой «*(ал)*» и имеет морфологический тег «*<v>*», означающий *глагол*.

Для тех пользователей, которым подобный текстовый способ ввода различных параметров поиска может показаться неудобным, разработан удобный графический режим, где грамматические теги можно выбирать проставляя галочки и система сама правильно оформит написание леммы, начала, середины и конца слова.

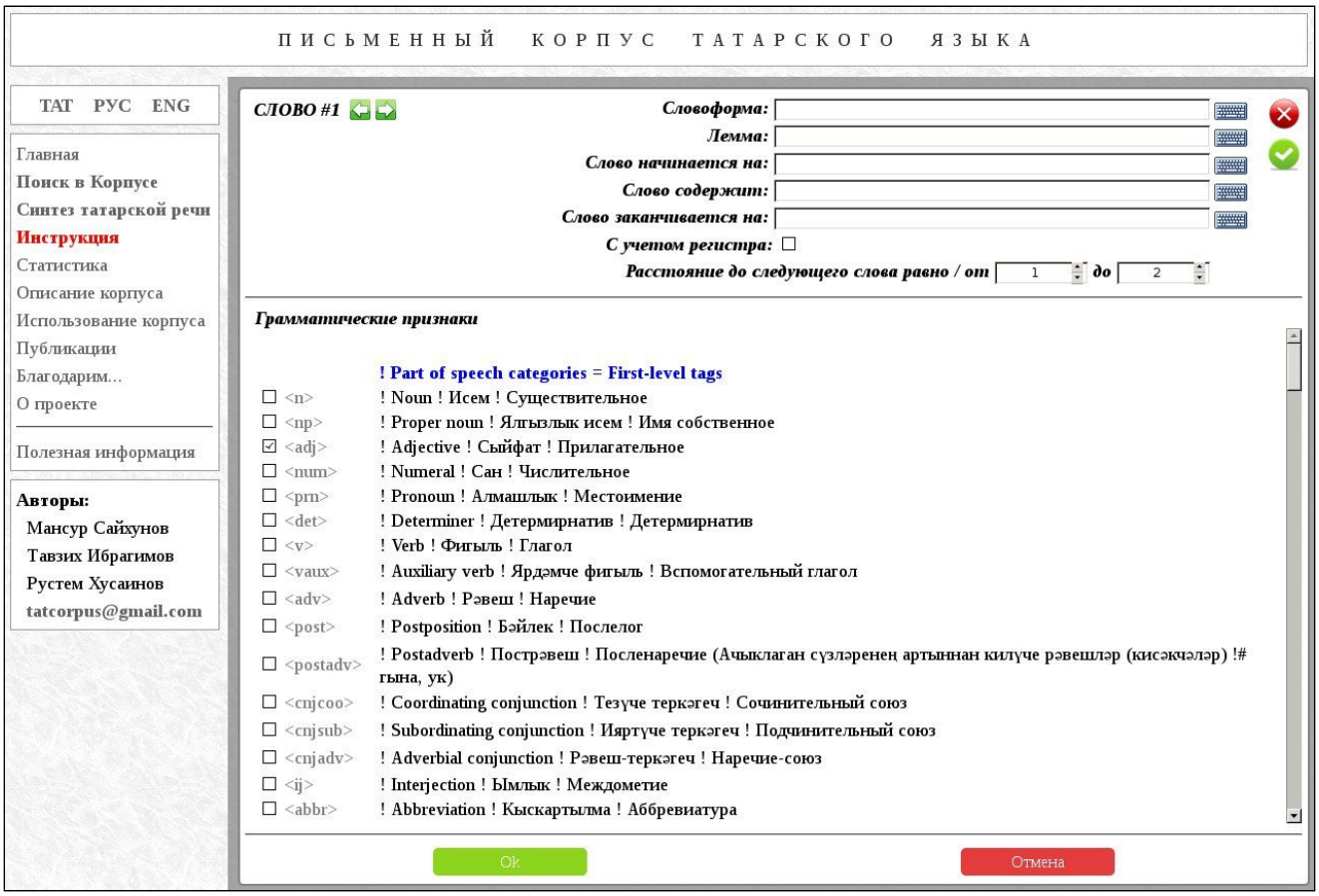

Рис. 3. Графический режим ввода параметров поиска

Руководство пользователя на русском, татарском и английском языках, описывающее все возможности Письменного корпуса татарского языка расположено на сайте в разделе «Инструкция» [2].

4. Техническое описание проекта

Система сложного морфологического поиска устроена следующим образом. В связи с тем, что необходимо обеспечить быстрый поиск по таким параметрам, как словоформа, лемма, морфологические теги, шаблоны, расстояния между словами, было принято решение написать серверную часть поиска на языке программирования **C**. Иными словами, данный модуль с рабочим названием **fastmorph** изначально задумывался как изолированная по отношению к веб-серверу система. Fastmorph на этапе инициализации загружает все необходимые данные корпуса из СУБД MySQL и компактно размещает их в виде массивов в оперативной памяти компьютера, что:

- позволяет избежать потери времени при операциях обращения к жесткому диску;
- делает архитектуру приложения максимально простой и гибкой для дальнейшего расширения функциональности или адаптации под другие проекты.

Реализованные на языке PHP дополнительные программы производят "трансляцию" относительно свободного запроса пользователя в строго типизированный набор данных и передают системе fastmorph. Fastmorph выполняет поиск в своей индексированной базе и возвращает список найденных предложений с метаинформацией (список источников, выделение в предложениях искомых элементов специальными тегами, указание лемм и набора тегов найденных элементов, общее количество найденных примеров и др.) обратно PHP модулю, а тот, в свою очередь, — пользователю.

Текущая идея заложенная в основу алгоритма позволяет эффективно утилизировать как кэш процессора, так и вычислительные мощности мультиядерных систем. В итоге, удалось добиться большой скорости выполнения поиска в 0,2-2 секунды в зависимости от параметров поиска даже на скромном оборудовании. При этом не используются такие ухищрения, как отображение примерного количества совпадений и отложенный (фоновый) поиск.

22 ноября 2016 года исходный код корпусного поискового движка **fastmorph** был открыт под лицензией GNU General Public License v3, что позволяет всем желающим использовать наши наработки в своих проектах.

Литература

1. *Сайхунов М. Р., Ибрагимов Т. И., Хусаинов Р. Р.* Письменный корпус татарского языка [Электронный ресурс]. Казань, 2012. Режим доступа: http://corpus.tatar, свободный.

2. *Сайхунов М. Р., Ибрагимов Т. И., Хусаинов Р. Р.* Письменный корпус татарского языка. Руководство пользователя [Электронный ресурс]. Казань, 2015. Режим доступа: http://corpus.tatar/manual.htm, свободный.

3. *Aksan, Y., Aksan, M.* Building a national corpus of Turkish: Design and implementation. Working Papers in Corpus-based Linguistics and Language Education no. 3, pp.299-310. Tokyo: TUFS, 2009.

4. Apertium - Открытая платформа машинного перевода [Электронный ресурс]. Режим доступа: http://wiki.apertium.org/wiki/Publications, свободный.

5. *Jurafsky, Daniel, and James H. Martin.* Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall, 2009. 1024 p.

6. *Kilgarriff, Adam.* Linguistic search engine. In Kiril Simov, editor, Shallow Processing of Large Corpora: Workshop held in association with Corpus Linguistics. Lancaster, 2003. Pp. 53-58.

COMPLEX MORPHOLOGICAL SEARCH SYSTEM FOR THE CORPUS OF WRITTEN TATAR LANGUAGE

M. R. Saykhunov, R. R. Khusainov, T. I. Ibragimov
Kazan

Annotation. This article describes features, creation history and main difficulties which autors have faced during their work on the system of complex morphological search for the Corpus of Written Tatar.

Keywords: corpus linguistics, Tatar language, natural language processing, fast search, morphology, Internet, programming.

Contact information: tatcorpus@gmail.com.