КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.912

ПОДХОД К СОЗДАНИЮ КОМПЛЕКСА ИНСТРУМЕНТОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Н. Е. Балакирев*, Е. В. Полицына**

**Московский авиационный институт (национальный исследовательский университет)*
*** Национальный исследовательский университет «Высшая школа экономики»*

Поступила в редакцию 05.03.2016 г.

Аннотация. Необходимость создания автоматических систем и сервисов для решения самых разных задач, отсутствие универсальных алгоритмов анализа текстов, разные требования к скорости работы инструментов при обработке больших объемов данных делают актуальной разработку комплекса инструментов автоматизированного анализа текстов с наличием набора сервисов для быстрого решения задач любым пользователем с любого устройства, инструментов для проведения более глубоких исследований текстов, разработки собственных инструментов и возможностью его использования как основы для дальнейших исследований в области компьютерной лингвистики и апробации новых алгоритмов анализа.

Предлагаемый комплекс включает в себя открытую систему автоматизированного анализа текстов, портал «Автоматизированный анализ текста», набор сервисов и мобильных приложений. Его использование обеспечит возможность создания и использования программных инструментов для решения практических задач, учитывая интересы широкого круга пользователей, а также способствует продвижению исследований в области компьютерной лингвистики и внедрению их результатов в прикладные программы.

Ключевые слова: инструменты автоматизированного анализа текстов, система автоматизированного анализа текстов, инструменты проведения исследований в области компьютерной лингвистики.

Annotation. The need of creating new automated systems and services for a wide variety of applications, the lack of universal algorithms for text analysis, different performance requirements of the large data volumes processing make it urgent to develop integrated tools of automated text analysis and deeper text study, quickly solve user problems on any device, and to create the basis for further research in the field of computational linguistics and testing of new algorithms of analysis.

The proposed toolset includes an open system of automated text analysis, web-portal «Automated text analysis,» a set of services and mobile applications. Its usage will provide an ability to create and use software tools to solve practical problems, taking into account the interests of a wide range of users, as well as it will help to promote the research in the field of computer linguistics and introduce their results into various applications.

Keywords: automated text analysis tools, system of automated text analysis, computer linguistics research utilities.

ВВЕДЕНИЕ

Задачи автоматизированного анализа текстов из области теоретических исследований и экспериментальных систем все больше переходят в практическое русло. На протяжении нескольких десятилетий множеством ученых были предложены различные подходы к решению задач практических задач автоматического и автоматизированного перевода текстов с одного естественного языка на другой, классификации и кластеризации, поиска и др., созданы программы, их реализующие [1–5]. В их основе лежат результаты многочисленных работ в области компьютерной лингвистики Ю. Д. Апресяна, М. Мински, И. А. Мельчука, Д. А. Поспелова, Р. Шенка, И. Уилкса, Г. Г. Белоногова, А. К. Жолковского, Ч. Филмора, Д. Джурафски, М. Г. Мальковского, В. А. Фомичева, Н. Н. Леонтьевой, А. В. Сокирко, А. А. Кретова и др., посвященных созданию способов представления знаний и их автоматическому построению, алгоритмам морфологического и синтаксического анализа текста, алгоритмам выделения различных статистических характеристик текстов, анализу больших объемов текстовых данных и т. д.

Возросшие объемы обрабатываемой людьми текстовой информации привели к необходимости применять автоматизированные средства для решения самых разных задач широким кругом пользователей, используя для этого как компьютеры, так и мобильные устройства. В настоящее время соотношение компьютеров и мобильных устройств сильно смещается в сторону последних, согласно прогнозам, эта тенденция будет и дальше продолжаться [6]. Результаты автоматизированного анализа текстов необходимы не только для решения научных задач и обработки большого объема данных, но и в повседневной деятельности в профессиональных и личных целях: в задачах не только поиска и перевода на другие языки, но и SEO-оптимизации веб-систем, быстрого определения темы и краткого содержания текста, проверки орфографии и пунктуации, резюмирования текста для его адаптации к соответствующей аудитории, ускорения восприятия или отображению на мобильном устройстве [7, 8].

Несмотря на многообразие подходов к решению задач автоматизированного анализа текста не существует универсального решения каждой из них для реализации их в программных комплексах и приложениях для широкого круга пользователей. В большинстве случаев это экспериментальные системы, достаточно широко распространенные системы поиска и машинного перевода или немногочисленные сервисы, созданные для решения отдельной задачи.

Существуют разработанные за рубежом архитектуры и наборы библиотек и инструментов для создания, исследования и использования широкого спектра различных моделей анализа, а также интеграции их с приложениями для решения практических задач. Для английского языка созданы системы, позволяющие комбинировать наборы предоставленных инструментов анализа текста, но они реализованы в виде набора библиотек для разработки программного обеспечения и практически не поддерживают работу с текстами на русском языке (GATE, UIMA, Apertium, LingPipe).

Необходимость создания автоматических систем и сервисов для решения самых разных задач, отсутствие универсальных алгоритмов анализа текстов, разные требования к скорости работы инструментов при обработке больших объемов данных делают актуальной разработку комплекса инструментов автоматизированного анализа текстов с наличием набора сервисов для быстрого решения задач любым пользователем с любого устройства, инструментов для проведения более глубоких исследований текстов, разработки собственных инструментов и возможностью его использования как основы для дальнейших исследований в области компьютерной лингвистики и апробации новых алгоритмов анализа.

СТРУКТУРА КОМПЛЕКСА

В основу предлагаемого комплекса инструментов положена *открытая система*

автоматизированной обработки текстов [9] на русском языке, которая включает в себя набор инструментов обработки текстов, накопления полученной информации и ее последующего анализа. Система включает в себя клиент-серверное многопользовательское веб-приложение, имеющее графический интерфейс пользователя для доступа к инструментам базовой и аналитической обработки и API для обеспечения возможности использования инструментов системы в других приложениях и систему управления данными и инструментами.

Структура предлагаемого программного комплекса представлена на рис. 1.

Открытая система автоматизированной обработки текстов в первую очередь предназначена для проведения разноплановых исследований в области компьютерной лингвистики, написания пользователями собственных алгоритмов, апробации новых алгоритмов для улучшения работы ядра системы и т. д. Она предоставляет наиболее разнообразный набор инструментов анализа и платформу для создания новых инструментов, на основе наиболее отлаженных из которых создаются новые функции API.

На базе системы создан ряд сервисов для решения задач классификации и получения статистических характеристик текстов, представленных *на портале «Автоматизированный анализ текста»* (textanalysis.ru). Созданные на портале сервисы предоставляют пользователю простой и понятный интерфейс для работы с открытой системой автоматизированной обработки текста.

Использование сервисов дает возможность решения практических задач, как используя имеющуюся базу системы анализа, так и расширяя ее для своих задач. По мере реализации новых алгоритмов и появления новых функций API в открытой системе автоматизированной обработки текстов возможно быстро осуществлять разработку новых сервисов. Все это способствует продвижению исследований в области компьютерной лингвистики и является первым шагом к созданию более удобных инструментов анализа текстов

Рис. 1. Структура комплекса инструментов автоматизированного анализа текстов на русском языке

и расширению их функционала и интерфейса для широкого круга пользователей.

В настоящее время огромное распространение имеют мобильные устройства, которые используются далеко не только для связи. Их мощность и набор программных средств позволяют решать многие задачи без наличия компьютера или ноутбука, что с одной стороны упрощает человеку жизнь, но с другой стороны заставляет постоянно оперировать огромным объемом разноплановой информации, чему еще больше способствует наличие доступа к сети Интернет. При этом восприятие информации с экрана мобильного устройства еще больше усложняет ее быстрое понимание.

Поэтому следующим шагом является создание новых *веб-сервисов и мобильных приложений* для решения повседневных задач, связанным с быстрым получением необходимой информации разных групп людей: кратких вариантов текстов из сети Интернет, новостей, информации из QR-кодов, темы и основных понятий текста, в том числе на другом языке и т. д.

ОТКРЫТАЯ СИСТЕМА АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТОВ

В основе системы лежат три направления обработки текста: лингвистическое, статистическое, аналитическое. Средой объединения является система хранения, которая предполагает хранение промежуточных и окончательных результатов и инструментов, предназначенных для анализа текста. Система хранения включает в себя базу данных и облачный сервис в виде файлового пространства пользователя. Инструменты лингвистической и статистической обработки представляют собой наборы операций, отдельных алгоритмов анализа и компонентов. Основой для принципов развития и наполнения системы является аналитическая система накопления поступающей информации, построенная в соответствии с моделью адаптивно-динамического преобразования информации [9, 10], это делает систему динамической и позволяет расширять внутренние «знания» системы, для представление которых необходимо создание и реализация работы с семантическими сетями и онтологиями, позволяющими максимально отражать особенности текстов, заключенную в них информацию, понятия, связи между ними. Основными источниками информации для пополнения «знаний» системы являются толковые словари и тезаурусы.

Все инструменты базовой обработки привязаны к основным этапам автоматизированного анализа текста: графематическому, морфологическому, синтаксическому, семантическому. На вход инструментам базовой обработки поступает загруженный пользователем текст, на выходе получаются результаты статистического и лингвистического анализа в виде структур различных типов: словники, списки предложений, семантические сети, частотные распределения букв, слов, связей между словами и т. д. Полученные структуры могут представлять собой как конечный результат анализа, так и использоваться в качестве входных данных для инструмента аналитической обработки – языка сценариев – расширяемого инструмента, ориентированного на максимальную простоту использования. В его основу положено использование операций, выполняемых над извлеченной информацией [10]: наборами слов с их морфологическими характеристиками, понятий, предложений, синтаксических структур и т. д., что позволяет решать сложные задачи на большем объеме исходного материала и обрабатывать конструкции исходных данных более высокого уровня сложности. Язык предназначен для написания пользователями собственных алгоритмов анализа, в системе также представлены типичные «шаблонные» алгоритмы, для которых возможна настройка параметров обработки. После выполнения сценариев получаются результаты анализа в виде новых структур или численных значений.

Система реализована средствами Java EE с использованием набора инструментов GWT, технологии Servlet и набора библиотек анализа текстов: библиотеки морфологического анализа текстов основанной на использовании словаря А. А. Зализняка [11], библиотек

статистического и лингвистического анализа текстов. В дальнейшем планируется расширение функционала системы как путем реализации новых инструментов, так и использования существующих сторонних библиотек, распространяемых по свободной лицензии, в том числе GATE, Apertium, Textocat, API Yandex и Google.

Таким образом, система дает возможность исследовать потребности пользователей и предоставляет им платформу для самостоятельного создания инструментов, не прибегая к использованию языков программирования. Это способствует выявлению актуальных направлений в области практического применения алгоритмов автоматизированного анализа текстов и реализации и отладки новых инструментов.

ПОРТАЛ «АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ТЕКСТА» И СЕРВИСЫ АНАЛИЗА ТЕКСТА

Портал «Автоматизированный анализ текста» был создан для получения более полной информации о теоретических основах компьютерной лингвистики и результатах исследований в этой области. Портал содержит информацию по основам компьютерной лингвистики, перечень популярных программ по обработке текстов, публикации и литературу по автоматизированному анализу текстов.

Путем выделения некоторых отработанных сценариев, решающих практические задачи, в самостоятельные сервисы анализа [12] на базе открытой системы автоматизированного анализа текстов создаются инструменты для их использования вне системы. Каждый сервис системы имеет API для его вызова, формат для передачи параметров запуска сценария анализа в сервисе и формат передачи полученного результата.

Отдельным разделом портала «Автоматизированный анализ текста» являются сервисы для решения простых задач, в основе работы которых лежит использование API открытой системы автоматизированной обработки текста. Разработанные сервисы анализа текста, позволяют проводить базовый анализ текста, используя инструменты статистической обработки текста на разных уровнях анализа [13]:

- графематический – статистика встречаемости в текстах или наборе материала отдельных букв в общем и в определенной позиции слова, максимальная длина слова; максимальная длина «нового» слова (слова, которого нет в словаре морфологии А. А. Зализняка); максимальное слово; максимальное «новое» слово; средняя длина слова; средняя длина предложения; количество предложений; частотное распределение длин слов и т.д.
- морфологический – количество слов в тексте; размер словника; частотное распределение слов; частотное распределение неслужебных частей речи.
- синтаксический – частотное распределение сочетаемости слов, частотное распределение структур.

В самостоятельные разделы вынесены сервис автоматической классификации текстов [13], который позволяет получать результаты классификации и предоставляет возможность настраивать сервис для работы с любыми предметными областями, и сервис получения ключевых слов загруженного текста [14]. Сервис автоматической классификации имеет два режима работы: классификации и обучения. Во второй версии сервиса была реализована поддержка возможности создания иерархического классификатора и привязки областей к номерам областей в УДК. Работа сервиса была апробирована путем разработки и внедрения классификатора в области управления проектами и анализе результатов последующей классификации текстов.В дальнейшем планируется загрузка классификаторов других предметных областей, обучение системы и применение разных алгоритмов классификации для улучшения качества работы сервиса.

Сервис выделения ключевых слов в настоящее время базируется на использовании реализованного на языке сценариев алгоритма, основанного на учете частотного распределения имен существительных. Для улучшения работы сервиса планируется его сочетание с методом маркемного анализа А. А. Кретова [15].

Сервисы позволяют работать с уже ранее загруженными текстами, используя один и тот же компьютер и браузер, или загрузить исследуемый текст. Сохранение загруженных текстов предоставляет возможность многократного обращения к сервису без необходимости повторной загрузки текстов на сервер. Кроме того, интеграция сервисов системы анализа с порталом позволяет использовать общее файловое пространство пользователя и легко переходить от быстрого использования сервисов к более глубокому анализу и созданию собственных алгоритмов извлечения информации из текстов и ее обработки непосредственно в системе.

Все это позволяет обеспечить интеграцию сервисов анализа текстов на портале с открытой системой автоматизированной обработки текстов и предоставить пользователям как возможность быстрого и удобного доступа к выделенным сервисам в автоматическом режиме, так и легкий переход к полному функционалу системы анализа с возможностью детальной настройки обработки и написания собственных алгоритмов автоматизированного анализа текста.

МОБИЛЬНЫЕ ПРИЛОЖЕНИЯ АНАЛИЗА ТЕКСТА

С увеличением количества смартфонов и планшетов растёт потребность и в различных приложениях для решения самых разных задач, так как их использование дает возможность получения постоянного доступа к большому объему данных, а доступ к сети Интернет позволяет в любой момент времени обновить их. Необходимость просматривать такое количество информации обуславливает актуальность создания приложений для мобильных устройств, которые позволят быстро получить тему текста, ключевые слова, краткое содержание и т. д., чтобы понять представляет ли эта информация какой-либо интерес.

Кроме того, в транспорте, в путешествиях и т. д. у человека в большинстве случаев есть доступ только к мобильному устройству, при том что потребность в получении необходимой информации может стоять остро. В таких случаях, большие требования предъявляются к скорости работы приложения и простоте его использования, т. к. решается задача – быстро получить представление об информации для принятия определенного решения, выраженной в текстовой форме, а не исследовать его.

Веб-сервисы и мобильные приложения перевода текста решают аналогичные задачи в части работы с иностранными языками. Несмотря на то, что качество автоматического перевода существенно уступает переводу человеком с использованием словаря, использование таких сервисов и приложений позволяет получить ответ намного быстрее с достаточным для решения проблемы уровнем качества.

По статистике операционная система Android является самой популярной ОС среди планшетов и мобильных телефонов [16], в ноябре 2015 года ОС Android занимала 61,59 % рынка, поэтому в первую очередь под эту ОС необходимо разрабатывать приложения анализа текстов, основываясь на использовании API открытой системы автоматизированной обработки текстов, Yandex API, Google API и др.

ЗАКЛЮЧЕНИЕ

Предлагаемый комплекс включает в себя открытую систему автоматизированного анализа текстов, портал «Автоматизированный анализ текста», набор сервисов и мобильных приложений. Использование предлагаемого подхода обеспечивает возможность создания и использования программных инструментов для решения практических задач, учитывая интересы широкого круга пользователей, а также способствует продвижению исследований в области компьютерной лингвистики и внедрению их результатов в прикладные программы.

СПИСОК ЛИТЕРАТУРЫ

1. *Мальковский М. Г.* Прикладное программное обеспечение: системы автоматической обработки текстов / Мальковский М. Г., Грацианова Т. Ю., Полякова И.Н. - М.: МГУ, издательский отдел факультета ВМК, 2000. – 52 с.

2. *Мальковский М. Г., Старостин А. С.* Алгоритм синтаксического анализа, используемый в системе морфологического анализа «TREETON» // Тр. междунар. конф. Диалог'2007. – М.Ж Изд-во РГГУ, 2007. – С. 516–524.

3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. – М. : МИЭМ, 2011. – 272 с.

4. Сайт рабочей группы АОТ. [Электронный ресурс] Режим доступа: www.aot.ru. (Дата обращения: 12.01.2016).

5. *Попов Э. В.* Общение с ЭВМ на естественном языке / Э. В. Попов. – М. : Едиториал УРСС. 2004. – 360с.

6. Schadler Ted Mobile Engagement Providers Will Be A New $32.4 Billion Market By 2018 [Электронный ресурс]. Режим доступа: http://www.forbes.com/sites/forrester/2013/08/09/mobile-engagement-providers-will-be-a-new-32-4-billion-market-by-2018/. (Дата обращения: 12.01.2016)

7. *Глухов В. П.* Основы психолингвистики: учеб. пособие для студентов педвузов. – М. : АСТ: Астрель, 2005. – 351,[1] с, — (Высшая школа).

8. *Фрумкина Р. М.* Психолингвистика / Р. М. Фрумкина. – М. : Академия, 2003. – 320 с.

9. *Балакирев Н. Е., Добрышина Е. В.* Концептуальная модель и структура системы обработки текстовой информации // Информационные технологии. – 2010. – № 2. – С. 2–7.

10. *Балакирев Н. Е., Полицына Е. В.* Язык сценариев как инструмент аналитической обработки в открытой системе автоматизированного анализа текста // Вестник ВГУ. – 2013. – № 1. – С. 162–168.

11. *Зализняк А. А.* Грамматический словарь русского языка. – М. : Русские словари, 2003.

12. *Полицына Е. В.* Создание настраиваемого сервиса классификации в составе открытой системы автоматизированного анализа текста // Материалы XIII Международной научно-методической конференции «Информатика: проблемы, методология, технологии». – Т. 1. – Воронеж, 2013. – С. 80–84.

13. *Громова С. Н., Мельник Е. П., Полицына Е. В.* Внедрение сервисов открытой системы автоматизированной обработки текста на портале «Автоматизированный анализ текста» // Материалы XIII Международной научно-методической конференции «Информатика: проблемы, методология, технологии». – Т. 1. – Воронеж, 2013. – С. 80–84.

14. *Бочарова Р. Р.* Разработка требований и инструментов обработки текстов на основе анализа потребностей пользователей // Научные труды Международной молодежной научной конференции «XLI Гагаринские чтения». – М. : МАТИ, Т. 4, 2015. – С. 149–150.

15. *Кретов А. А.* Функциональный подход к выделению ключевых слов: методика и реализация / А. А. Кретов, И. Е. Воронина, И. В. Попова, Л. В. Дудкина // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2009. – № 1. – С. 68–72.

16. Top 8 Mobile&TabletOperatingSystem sonNov 2015 | StatCounterGlobalStats [Электронный ресурс] / StatCounterGlobalStats. – Режим доступа: http://gs.statcounter.com/#mobile+tablet-os-ww-monthly-201511-201511-bar. – (Датаобращения: 12.01.2016)

Балакирев Николай Евгеньевич – к.т.н., профессор кафедры «Проектирование вычислительных комплексов» НИУ «Московский авиационный институт».
Тел.: 8(499)141-94-82
E-mail: balakirev1949@yandex.ru

Полицына Екатерина Валерьевна – к.т.н., доцент департамента программной инженерии факультета компьютерных наук, Национальный исследовательский университет «Высшая школа экономики».
Тел.: 8(903)524-78-25
E-mail: kathrin.beaver@mail.ru

Balakirev Nikolay E. – candidate of technical sciences, professor, department of «Design of computing systems», Moscow Aviation Institute (National Research University).
Tel.: 8(499)141-94-82
E-mail: balakirev1949@yandex.ru

Politsyna Ekaterina V. – candidate of technical sciences, associate professor, faculty of computer science, school of software engineering, National Research University Higher School of Economics.
Tel.: 8(903)524-78-25
E-mail: kathrin.beaver@mail.ru