УДК 004.912

М. Н. Мокроусов, кандидат технических наук
ИжГТУ имени М. Т. Калашникова

АВТОМАТИЗИРОВАННАЯ СИСТЕМА НОРМАЛИЗАЦИИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ*

В статье представлена автоматизированная система нормализации текста, разделяющая текст на слова, предложения и абзацы и выделяющая в тексте имена собственные, аббревиатуры и буквенно-цифровые последовательности символов. Приводится структура системы, описание ключевых моментов ее работы и результаты экспериментов.

Ключевые слова: автоматическая обработка текста, морфологический анализ, стемминг, нормализация текста, регулярные выражения.

В классических методах автоматической обработки текстов (АОТ) выделяют три крупных этапа анализа: морфологический, синтаксический и семантический. Для проведения этих этапов необходимо подготовить текст: разбить его на структурные фрагменты и назначить для каждого фрагмента некоторое формальное описание [1].

В литературе подготовка текста для автоматического анализа называется предобработкой, сегментацией, токенизацией, графематическим анализом, лемматизацией и т. п. Наиболее полно такая подготовка текста описана в [2] и решает следующие задачи: разделение входного текста на слова, разделители и т. д.; сборка слов, написанных в разрядку; выделение устойчивых оборотов, не имеющих словоизменительных вариантов; выделение имен собственных; выделение электронных адресов, имен файлов, дат, номеров телефонов, числительных; выделение предложений из входного текста; выделение абзацев, заголовков, примечаний.

Данный этап характерен для любого естественно-языкового лингвистического процессора и выполняется до этапа морфологического анализа, на котором для каждого слова определяется принадлежность к части речи, определяются его словоизменительная парадигма и грамматические признаки.

Задача нормализации особенно остро стоит во флективных языках с развитой омонимией, когда одно и то же слово может иметь несколько грамматических характеристик и относиться к разным классам графем и понятий.

На этапе нормализации выделяются следующие классы графем: слово, иностранное слово, буквенная последовательность (последовательность букв, которая не является словом), сокращение, целое число, дробное число, буквенно-цифровая последовательность, разделитель предложений, знак пунктуации.

Также при анализе текста необходимо учитывать такие атрибуты графем, как регистр (нижний, верхний, заголовочный – первый символ слова находится в верхнем регистре, смешанный – любая другая комбинация регистров), признак начала предложения, признак имени собственного (если слово присутствует в справочнике имен собственных), наличие предшествующего пробела. Атрибут предшествующего пробела присваивается графеме, если перед ней находился символ пробела. Данный атрибут используется для разрешения ряда неоднозначностей, в частности, связанных с использованием точки: наличие пробела может влиять на интерпретацию точки как разделителя предложений или часть другой графемы.

Фрагментация текста на предложения и лексемы осуществляется с помощью различного рода разделителей, которые присутствуют в тексте. К разделителям лексем относятся «()», «|», «{}», «[]», «<>», «/»,«;», «.», «,», «!», «?», «\», «:», «–», а также различные виды кавычек и пробельные символы (пробел, табуляция, символы конца строки).

Критериями конца предложения могут быть восклицательный и вопросительный знаки, стоящие в конце лексемы или выделенные в отдельную лексему, и точка, стоящая в конце лексемы или являющаяся самостоятельной лексемой, после которой следует текст с заглавной буквы. Также необходимо учитывать общепринятые сокращения, инициалы и т. п., точки, в конце которых нужно анализировать по отдельным правилам.

Правила выделения графем, по которым невозможно составить словари, используют механизм регулярных выражений – формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов [3]. Например, регулярное выражения для поиска денежной суммы в долларах будет иметь вид *\$[0-9]+(\.[0-9][0-9])?*, а для поиска e-mail: *[a-z0-0][a-z0-9._-]+@[a-z0-9._-]+\.[a-z0-9]+*.

Для приведения слова к нормальной форме использовался грамматический словарь Зализняка, программно реализованный в библиотеке функций mcr.dll, и программа морфоанализа компании Yandex – mystem, которая позволяет «угадывать» по морфемному составу слова его грамматические характеристики и лемму.

В настоящее время в существующих системах АОТ модуль нормализации используется и как самостоятельная подсистема для выделения именованных сущностей и фактов, и в составе комплексного анализатора, результаты которого используются для информационного поиска, перевода с одного языка на другой, построения реферата, сравнении текстов и т. д. Такими системами являются: Арион-Лингво, TextAnalyst, ЭТАП3, RCO Fact Extractor SDK и мно-

* Работа выполнена при частичной поддержке Госзадания МОиН РФ Проект № 625 и гранта РФФИ № 15-07-08077.

гие другие. Все они представляют собой готовые решения с закрытыми базами данных и правилами их обработки.

В предлагаемой системе модуль пополнения правил и данных открыт для редактирования. На рис. 1 представлена структура такой системы.

Система управления нормализацией текста выполняет функции общей координации работы системы. В этот модуль поступают все действия пользователя, а также уведомления о внутренних событиях системы. В соответствии с полученной информацией этот модуль активизирует другие подсистемы для выполнения необходимых операций.

Модуль анализа текста является ядром системы, выполняя функции по разделению входного текста на слова и предложения. Данный модуль взаимодействует с модулями проверки слова по маске и словарю и выделения найденного слова.

Модуль взаимодействия с базой данных инициализирует базу данных и предоставляет интерфейсы для доступа к таблицам посредством технологии ADO. Также данный модуль обслуживает выполнение запросов к базе данных и получает ответы от нее.

Модуль визуализации результатов обеспечивает понятное конечному пользователю отображение результатов в табличной форме с использованием информационных тегов и цветовым выделением фрагментов текста.

Модуль поиска слова по маске, словарю выполняет функции выделения имен собственных и таких понятий, как электронные адреса, имена файлов, числовые значения и т. д. В данном модуле используется механизм регулярных выражений.

Модуль поиска лексемы осуществляет приведение слова к ее нормальной (словарной) форме – лемме. Например, в русском языке для существительных нормальной формой считается именительный падеж, единственное число. Данный модуль необходим для поиска по словарю, т. к. в словаре содержатся только нормальные формы слова. Также данный модуль позволяет выявить возможные грамматические характеристики слова, которые будут уточнены на этапах морфологического и синтаксического анализа.

Модуль сегментации необходим для выделения слов в тексте соответствующими информационными тегами или цветом.

Модуль экспорта/импорта текста обеспечивает возможность загрузки исходного текста из файла, а также сохранения обработанного текста с тегами нормализации.

Модуль работы с регулярными выражениями обеспечивает поддержку регулярных выражений в алгоритмах поиска. Для работы с регулярными выражениями используется открытая библиотека TRegExpr, написанная для Delphi.

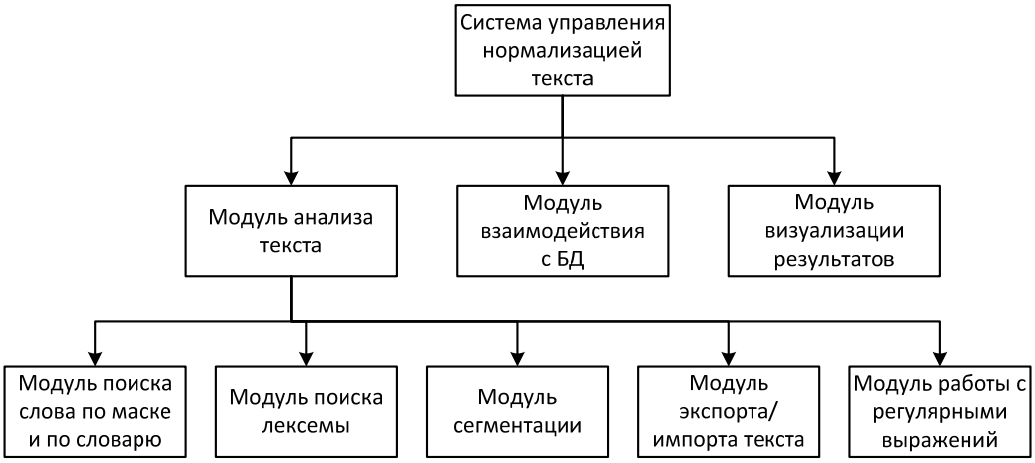


Рис. 1. Структура системы нормализации текста

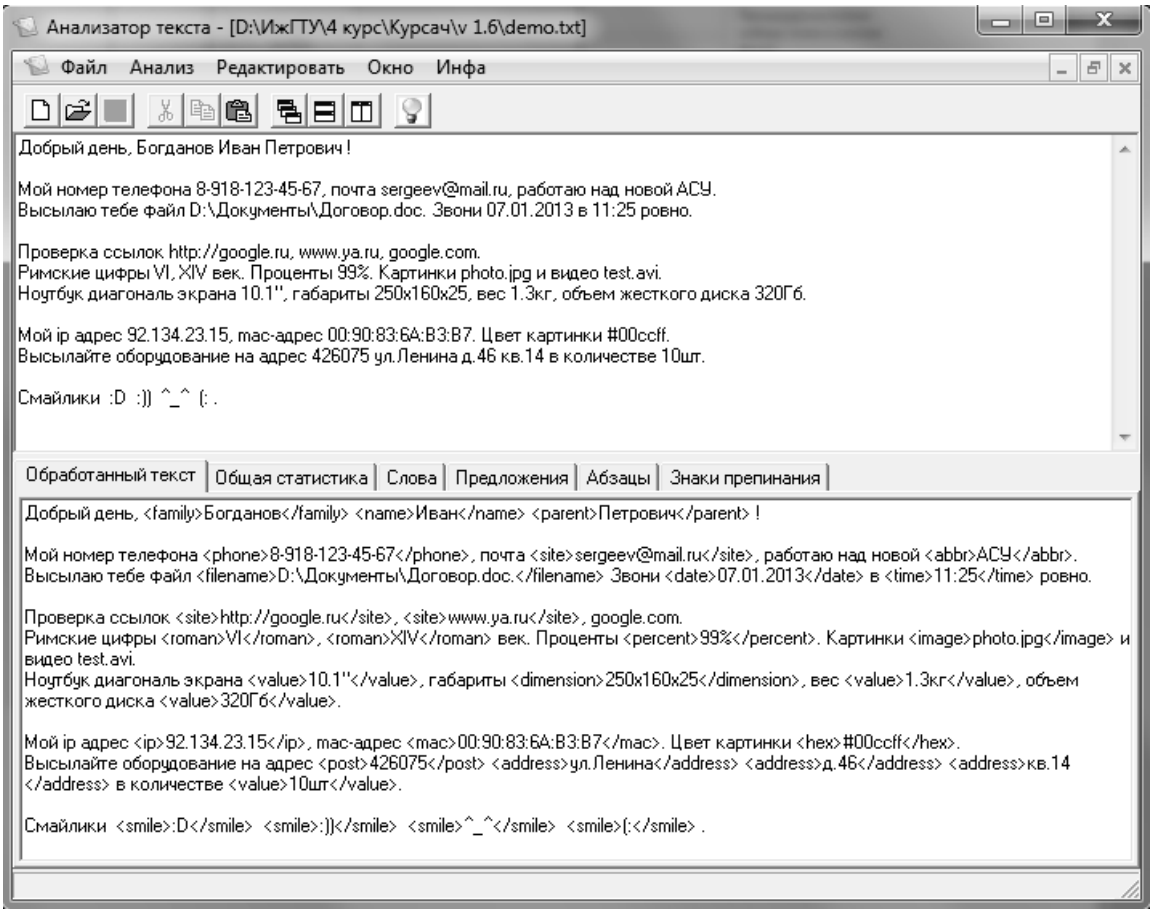
В ходе работы программы исходный текст сегментируется на слова, предложения и абзацы. Найденные в словаре или по шаблону графемы выделяются тегами или цветом, строятся таблицы с указанием количества найденных графем. На рис. 2 показан интерфейс разработанной системы, где видно выделение найденных графем тегами (рис. 2, *а*) и содержимое некоторых статистических таблиц. Так, на рис. 2, *б* показана вкладка «Слова», где отображаются порядковый номер и позиция графемы в тексте, ее нормальная форма, грамматическое и функциональное описание. На рис. 2, *в* показана «Общая статистика», в которой отображено количество найденных графем каждого класса.

В настоящее время база данных системы содержит 932334 имени собственных разбитых на 18 классов, 5357 аббревиатур из 89 предметных областей и 69 регулярных выражений для 21 буквенно-цифровой графемы.

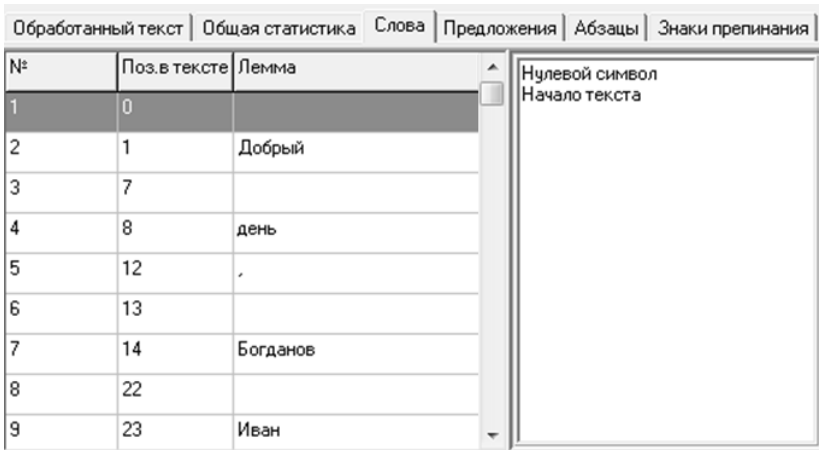
Для проведения экспериментов с системой были выбраны фрагмент из романа «Анна Каренина» Льва Толстого (249 слов, 1669 знаков), отрывок из учебного пособия «Физика. Классический курс»* (270 слов, 1861) и отрывок из крымской речи Владимира Путина от 18 марта 2014 года (247 слов, 1773 знака). В результате анализа указанных фраг-

* Мякишев Г. Я., Буховцев Б. Б., Чаругин В. М. Физика. 11 класс. Учебник. 19-е изд. М. : Просвещение, 2012. 400 с.

ментов из 74 имен собственных было найдено 56 (75,67 %), из 26 буквенно-цифровых графем было найдено 24 (92,30 %), из 13 аббревиатур было найдено 8 (61,53 %). В некоторых случаях программа считала именами собственными повествовательные имена, т. к. они присутствовали в справочнике, а некоторые аббревиатуры были привязаны к цифровым графемам и учитывались как буквенно-цифровая последовательность.



а



б

в

Рис. 2. Экранные формы приложения: *а* – главная форма программы; *б* – вкладка «Слова»; *в* – вкладка «Общая статистика»

Таким образом, представленная система нормализации текста может быть использована как самостоятельная программа для выделения в тексте именованных сущностей и нелексических буквенно-цифровых фрагментов текста, так и в качестве модуля предобработки текста для задач АОТ. Расширяемая база данных имен собственных и правил поиска на основе регулярных выражений делает систему гибкой для применения в разных предметных областях.

Библиографические ссылки

1. *Мокроусов М. Н.* Интеллектуальный поиск в задаче извлечения знаний из естественно-языковых текстов // Всероссийская конференция с элементами научной школы для молодежи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации». – В 4 т. Т. 2. – Ульяновск : УлГТУ, 2009. – С. 347–355.

2. *Сокирко А. В.* Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) : дис. ... канд. техн. наук. – М., 2001. – 100 с.

3. *Гойвертс Я., Левитан С.* Регулярные выражения. Сборник рецептов. – СПб. : Символ-Плюс, 2010.

* * *

Mokrousov M. N., PhD in Engineering, Associate Professor, Kalashnikov ISTU

Computer-aided system of natural-language texts normalization

The article represents the computer-aided system of text normalization that divides the text into words, sentences and paragraphs and extracts the personal names, abbreviations and alphanumeric sequences of the symbols in the text. The structure of the system, the key points of its work and the results of experiments are described.

Keywords: natural language processing, text segmentation, stemming, text normalization, regular expressions.

Получено: 11.11.15