OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

**УДК 004.822:514**

# АВТОМАТИЧЕСКАЯ СИСТЕМА ДЛЯ ВЫЯВЛЕНИЯ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ И СВЯЗЕЙ ИЗ АНГЛОЯЗЫЧНЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

М.В.Макарич (*2348843@tut.by*)
*Белорусский государственный лингвистический университет,*
*г.Минск, Республика Беларусь*

В статье рассмотрен вопрос извлечения, систематизации и оперативного анализа текстовой информации. Перечислены основные проблемы, с которыми сталкивается пользователь при выделении информационных объектов из сети Интернет. Предложена автоматическая система для выявления информационных объектов и их связей из англоязычных публицистических текстов, репрезентующая их в виде, удобном для дальнейшей обработки и создания сбственных баз данных.

*Ключевые слова:* автоматическая система обработки текста, база данных, извлечение знаний, информационный объект.

Развитие современных технологий, связанных с хранением, обработкой и передачей текстовых данных повлияло на быстрое увеличение объемов разноплановой информации, представленной в сети Интернет. От способов ее извлечения, систематизации и оперативного анализа зависит успех деятельности не только в науке, технике и в образовании, но и в области материального производства, в социально-экономической и политической сферах [1, с.56]. До 85% новых знаний аналитики до сих пор получают, изучая тексты, поэтому в ближайшем будущем наиболее востребованными станут системы с максимально автоматизированными процессами структурирования содержания текста (extract, transfer, load — «извлечение, преобразование, загрузка») [2]. Важной чертой таких систем будет функция оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемой с помощью методов интеллектуального анализа текста.

Современную жизнь уже невозможно представить без Интернет-среды. Однако, при поиске в ней информации во многих случаях пользователям выдается огромное количество текстов, которые приходится фильтровать, т.е. выбирать из них нужную информацию. В то же время большинство пользователей интересуют лишь конкретные данные. Например, специалиста по кадрам интересуют организации, где человек работал, кем и когда это было. Средства массовой информации (СМИ) интересуют данные о передвижениях влиятельных лиц и политических деятелей, катастрофах, террористических актах и др. В последние десятилетия пользователи все чаще ищут информацию не только на русском, но и на других языках, чаще всего — на английском. При этом результаты должны быть представлены в требуемом виде, например, в тех формах, к которым привык пользователь или которые приняты в соответствующих службах, заинтересованных в получении

информации. Другой вариант – в формах, удобных в плане последующей обработки, например, в форме таблиц.

Например, важная задача многих служб безопасности – анализ потока информации (СМИ, сообщений специализированных агентств и др.) с перемещением политических деятелей, выявлением террористических групп, их деятельности, оценка степени их активности, способов борьбы с ними. Причем события, в которых участвуют группы или личности, должны быть привязаны к месту и времени. Отметим, что общественно-политическая деятельность во многих случаях носит международный характер. Отсюда необходимость работы с англоязычными текстами.

В то же время поисковые сервера сети Интернет предоставляют пользователям весьма краткосрочный срез информации, а использование рубрик, таких, как "политика", "культура", "экономика" и т. п. дают лишь поверхностное представление о предметной области и, как следствие, требуют практически "ручной" переработки десятков тысяч сообщений. При обращении к специализированным отечественным и зарубежным "новостным" серверам исследователь сталкивается с рядом проблем [3, с.11]:

- количество источников, необходимых для получения объективного представления об исследуемом объекте, значительно возрастает (от 100 до 1000 в расчете на одно исследование);
- в свободном доступе оказывается только информация "текущего дня", доступ к информационным архивам, содержащим нужную информацию о прошедших событиях, сопряжен со значительными финансовыми издержками;
- адресное, прямое обращение к зарубежным источникам может вызвать интерес иностранных спецслужб или конкурирующих организаций и, как следствие, источник информации может превратиться в источник дезинформации.

Поэтому для обеспечения информационной поддержки процесса исследования по интересующей проблеме остается актуальным вопрос создания и накопления собственного информационного ресурса, содержащего только интересующую пользователя конкретную информацию. Такую информацию называют — *информационными объектами* [4, с.236]. Отсюда возникает проблема выявления информационных объектов и их связей из текстов естественного языка. Для решения поставленной задачи нами создана автоматическая система для формализации содержания англоязычных публицистических текстов. Система TRT обрабатывает любой неструктурированный текст и представляет содержащуюся в нем информацию в виде таблицы.

Лингвистическая база данных системы TRT включает в себя лексико-семантический и семантико-синтаксический блоки. На первом этапе производится лексико-семантический анализ всех словоформ входного текста с присвоением им специальных семантических кодов в соответствии с разработанной нами классификацией для текстов узкой предметной области [5, с.79]. Далее, производится семантико-синтаксический анализ, включающий в себя сегментацию предложений обрабатываемого текста и установление связей между словами внутри выделенных синтаксических групп [6]. Результаты анализа заносятся в соответствующие графы таблицы, наглядно репрезентующей: *Кто действует? По отношению к кому (чему) действует? Как, каким образом действует? Где действует? Когда, в какое время действует?*

Программа системы TRT реализована на языке программирования C# (Sharp). Структура окна программы TRT в обычном режиме приведена на рисунке 1.

Как это видно на рисунке, основными элементами окна являются четыре рабочие области для записи, разработки и презентации конечного продукта системы – табличного реферата. Нижняя область содержит алфавитный словарь, снабженный семантическими

кодами. В левой верхней области находится обрабатываемый текст в txt. формате, в правой – массивы ключевых слов данного текста различной степени значимости. Работа с элементами окна производится с помощью выбора необходимого файла в строке заголовка. В правой нижней части окна расположена кнопка «GENERATE», с использованием которой программа переходит в режим создания табличного реферата текста, находящегося в данный момент в левой верхней области окна. После команды «GENERATE» в центре экрана появляется табличный реферат текста.

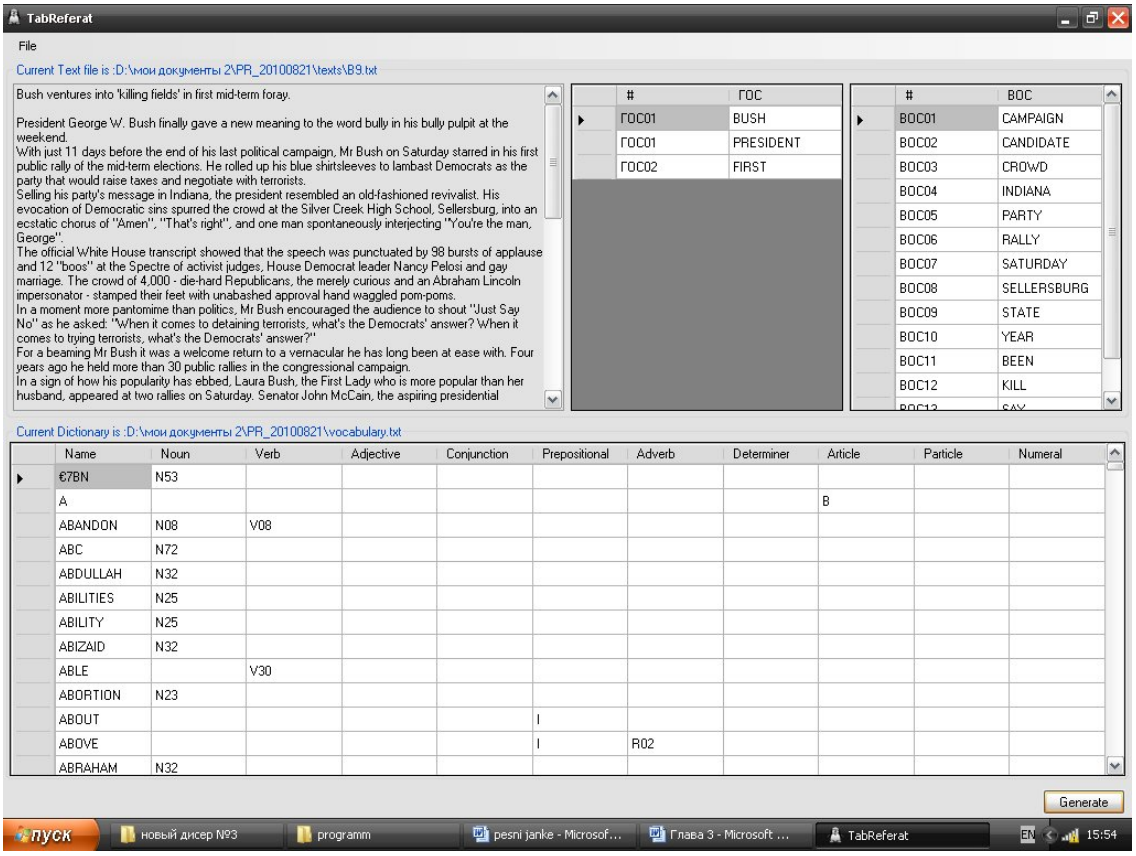


Рисунок 1 – Структура окна TRT в обычном режиме

Разработанная система автоматического реферирования TRT может быть использована для получения совокупности информационных объектов и их связей, извлеченных из массива газетно-публицистических текстов.

С использованием разработанной системы возможно создание баз данных с оперативным отслеживанием информации определенных субъектов по месту, времени и цели. В зависимости от конкретного предмета, интересующего пользователя, тексты могут быть подобраны с учетом определенной более узкой тематики.

Принимая во внимание тот факт, что в среду разработки системы может быть внесено неограниченное количество нового лингвистического материала из аналогичной или любой другой предметной области, можно рекомендовать использование данной системы для отбора существенных сведений из текстов любой предметной области.

На рисунке 2 приведена структура окна TR в рабочем режиме, содержащая результат работы данной системы – табличный реферат текста.

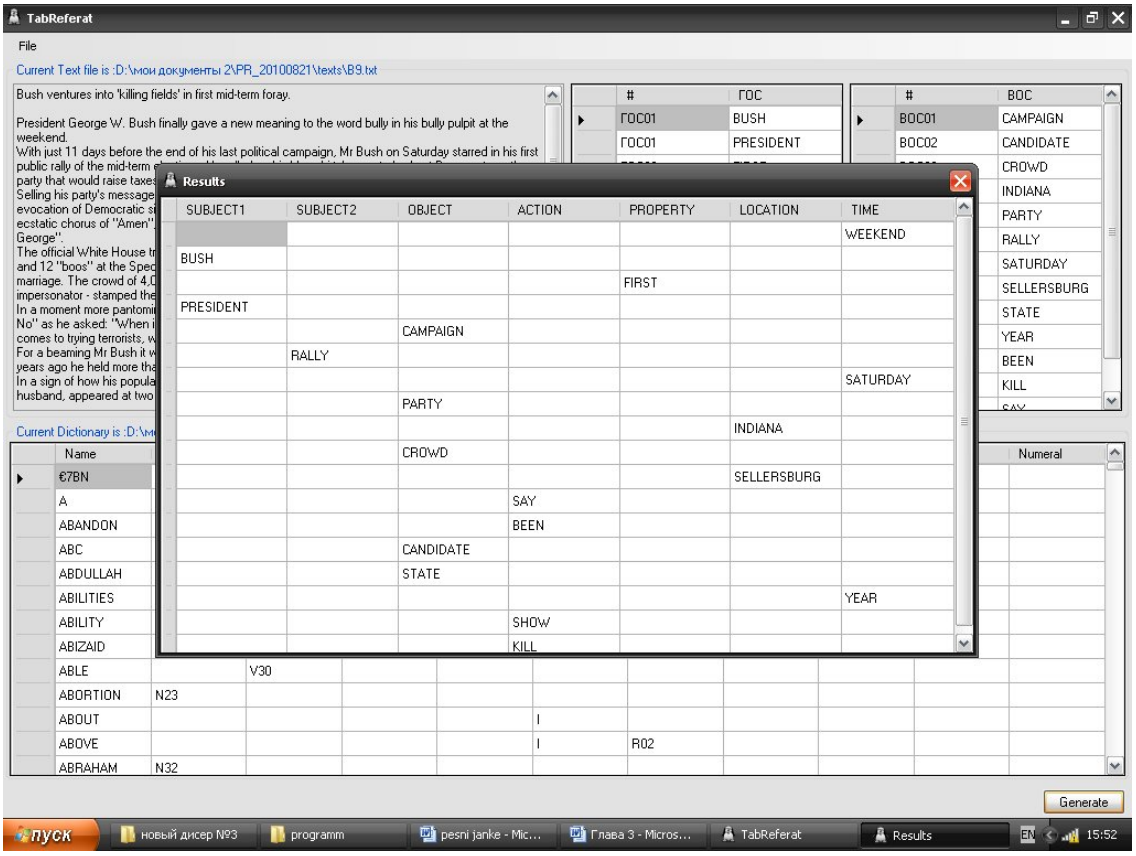


Рисунок 2 – Структура окна TRT в рабочем режиме

## Библиографический список


1. Васенин, В.А. и др. Автоматизированный анализ текстовой информации / В.А. Васенин [и др.] // Информационные технологии. – 2009. – №7. – С.56-57.
2. Рябышкин, В. Технологии извлечения знаний из текста // В.Рябышкин, Н.Ильин [и др.] // Открытые системы. – 2006. – №6. [Электронный ресурс]. – Режим доступа: http://www.osp.ru/os/2006/06/2700556/– Дата доступа: 11.08.10.
3. Кретов, В.С. Методика подготовки информационного ресурса для проведения политологических исследований на основе информации сети Интернет / В.С.Кретов, М.Н.Котов // 155М 0548-0027 НТИ.СЕР.2. Информ.процессы и системы. – 2006. – № 5. – С.11-15.
4. Кузнецов, И.П. Англо-русская система извлечения знаний из потоков информации в Интернет-среде / И.П.Кузнецов, Н.В.Сомин // Системы и средства информатики. Вып.17 / отв. Ред. И.А.Соколов. – М.: Наука. – 2007. – С.236-253.
5. Зубов, А.В. и др. Автоматический синтаксический анализ английского предложения как основа для формализации содержания текста / А.В.Зубов, М.В.Макарич // Прикладная лингвистика в науке и образовании: лингвистические технологии и инновационная образовательная среда: Коллективная монография. – СПб.: «ЛЕМА», 2010. – С. 76-84.
6. Макарич, М.В. Алгоритм автоматического синтаксического анализа англоязычного предложения / М.В.Макарич // Вестн. Минск.гос.лингв.ун-та. Сер.1, Филология. – 2010. – №3(46) – С.163-172.