КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.912

ЯЗЫК СЦЕНАРИЕВ КАК ИНСТРУМЕНТ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ В ОТКРЫТОЙ СИСТЕМЕ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТА

Н. Е. Балакирев, Е. В. Полицына

МАТИ – Российский государственный технологический университет имени К. Э. Циолковского

Поступила в редакцию 8.02.2013 г.

Аннотация. Обосновывается актуальность создания пользовательских инструментов анализа текстов, обеспечивающих открытость системы для расширения и модификации пользователями. Описываются особенности открытой системы анализа текста и заложенные в ее основу подходы к анализу текста. Предлагается язык сценариев как расширяемый инструмент анализа текста, предназначенный для широкого круга пользователей. Приводятся примеры реализованных алгоритмов для решения практических задач.

Ключевые слова: инструменты анализа текста, автоматизированный анализ текста, система анализа, обработка естественноязыкового текста, извлечение знаний, формализация естественного языка.

Annotation. The article reveals the necessity of creating new user-level text analysis tools which should provide facilities for the open text analysis system for extending its functionality by users. The article shows details of the open text analysis system and used text analyses approaches which it is based on. A script language is suggested as an expandable tool for text analysis for wide range of users. Some examples of the usage of this language are shown in scripts for some practical tasks.

Keywords: text analysis tools, automated text analysis, system of analysis, natural language text processing, data mining, formalization of natural language.

ВВЕДЕНИЕ

Необходимость создания открытой системы автоматизированного анализа текста обусловлена двумя причинами.

Потребность в обработке больших объемов информации, развитие программных и аппаратных средств и попытки создания интеллектуальных систем, связанных с проблемами «понимания» текста [1], привели к развитию методов автоматизированного анализа текста. Проведенный анализ показал, что большинство систем анализа текста носят экспериментальный характер, работают с ограниченным объемом информации или решают определенный круг задач, для расширения которого требуется внутренняя перестройка системы. Существующие системы не предоставляют гибких инструментов и возможностей изменения алгоритмов обработки текстов, для этого, как правило, предоставляются только наборы библиотек (GATE, UIMA, LingPipe), использование которых требует знаний в области программирования.

В связи с высокой сложностью задачи для развития систем автоматизированного анализа текста требуется участие множества специалистов из различных областей знаний, для чего необходимо наличие системы, ориентированной на широкий круг пользователей.

Это определяет актуальность как теоретических работ по созданию новых подходов к построению систем автоматизированного анализа текста, открытых для расширения и модификации пользователями, так и практических разработок пользовательских инструментов, реализующих в рамках этих систем функции обработки текста.

ОСОБЕННОСТИ ОТКРЫТОЙ СИСТЕМЫ АНАЛИЗА ТЕКСТА

В данном случае под открытостью понимается возможность изменения и расширения функционала системы пользователем путем использования набора инструментов, которые позволяют оценивать связь между алгоритмом и результатами, получаемыми при его выполнении.

Открытая система автоматизированной обработки текста представляет собой набор общедоступных инструментальных средств для анализа текстов. Основными особенностями системы являются:

- выделение уровней обработки текста;
- предоставление пользователю файлового пространства на сервере для хранения информации и результатов анализа;
- разделение процесса анализа текста на два уровня: базовую и аналитическую обработку;
- реализация инструментов базовой обработки внутри системы и их предоставление пользователям в качестве сервиса на специализированном портале;
- реализация инструментов аналитической обработки в виде языка сценариев для их легкой адаптации под различные прикладные задачи.

Открытость системы для пользователя обеспечивается:

- наличием гибких инструментальных средств анализа текста;
- учетом и возможностью использования извлекаемой в процессе анализа информации;
- возможностью расширения за счет включения в ее состав собственных методов и инструментов пользователя.

Открытая система анализа текста позволяет решать широкий набор исследовательских и практических задач и накапливать знания о естественном языке, способствуя решению задачи понимания текста.

ПОДХОДЫ К АНАЛИЗУ ТЕКСТА

Естественно-языковой текст воспринимается человеком исходя из его знаний, опыта и используемых методов анализа, накапливаемых в процессе практической деятельности, что позволяет глубже понимать смысл текста и заключенные в нем закономерности. Аналогично, открытость системы анализа текстов позволяет накапливать получаемые «знания» внутри системы и расширять набор инструментов и алгоритмов анализа, тем самым сохраняя получаемый опыт и методы обработки текста.

Решение задач автоматизированного анализа текста в общем случае включает в себя извлечение различной, но схожей по структуре информации и ее последующий более глубокий анализ, алгоритм и инструменты проведения которого зависят от конкретной цели [2]. Существующие системы анализа, прежде всего, включают в себя полный или сокращенный набор этапов обработки, принятых в компьютерной лингвистике: графематический, морфологический, поверхностный синтаксический, глубинный синтаксический, поверхностный семантический, глубинный семантический. Решение практических задач по анализу естественноязыкового текста не ограничивается получением информации на вышеперечисленных этапах, а требуют дополнительного анализа полученной информации.

Кроме этого, известно, что задача построения формальной модели отображения текста не имеет однозначного и универсального решения, что обусловлено сложностью естественного языка [3]. Поэтому необходимы гибкие инструменты, обеспечивающие множество подходов, которые бы позволили получать не только результаты анализа текста, но и строить различные алгоритмы их обработки. В качестве такого инструмента создан язык сценариев, отвечающий этим требованиям.

При исследовании текстов могут применяться три подхода:

1. Множество текстов – один алгоритм.
2. Один текст – множество алгоритмов.
3. Множество текстов – множество алгоритмов.

Создание и использование множества алгоритмов позволяет глубоко исследовать текст, рассматривать его с различных точек зрения, выявлять различные наборы его свойств. Применение их к большому набору текстов обеспечивает возможность накопления и обобщения, как актуальной информации, так и алгоритмов анализа, что способствует развитию исследований в области автоматизированной обработки текстов.

КРАТКОЕ ОПИСАНИЕ ЯЗЫКА СЦЕНАРИЕВ

Язык сценариев разрабатывался как расширяемый инструмент, ориентированный на мак-

симальную простоту использования. Он является составной частью системы анализа текста [4] и базируется на использовании операций, выполняемых над извлеченной информацией [5]: 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. Каждый сценарий соответствует определенной модели отображения набора свойств текста.

Использование языка сценариев позволяет обеспечить:

1. Открытость процесса анализа текста.

2. Возможность самостоятельной разработки и применения алгоритмов решения различных задач пользователем.

3. Легкость отладки и внесения изменений в созданные сценарии.

4. Возможность сохранения и неоднократного применения созданных сценариев.

Язык сценариев включает в себя операции двух видов: операции над структурами извлеченных данных и операции управления. В последующем планируется введение дополнительных операций и применение существующих к новым типам структур.

Действие каждой операции языка сценариев с одной стороны определяется структурой конструкций множества, к которому она применяется, с другой – параметром, который при этом учитывается (например, учет части речи, частоты, веса понятий и т.д.). В настоящее время выделяются следующие типы структур:

- словники;
- списки связей слов;
- предложения;
- семантические сети.

В последующем планируется введение дополнительных операций и применение существующих к новым типам структур.

По аналогии с теоретико-множественными операциями и с учетом особенностей естественно-языковых текстов и извлекаемой из них информации вводятся следующие **операции над структурами** данных, получаемыми на базовых этапах обработки:

1. *Объединение (Тип данных, Структура 1, Структура 2, [Параметры])* – объединение двух структур с учетом заданного набора параметров.

2. *Пересечение (Тип данных, Структура 1, Структура 2, [Параметры])* – пересечение двух структур с учетом заданного набора параметров.

3. *Разность (Тип данных, Структура 1, Структура 2, [Параметры])* – разность двух структур с учетом заданного набора параметров.

4. *Отношение (Тип данных, Структура 1, Структура 2, [Параметры])* – доля структур первого текста, присутствующих или не присутствующих во втором.

5. *Объединение с отсечением (Тип данных, Структура 1, Структура 2, [Параметры])* – объединение двух структур с учетом заданного набора параметров с последующим отсечением элементов по выбранному критерию.

6. *Удаление (Тип данных, Структура 1, [Параметры])* – удаление из структуры элементов с учетом заданного набора параметров.

7. *Выборка (Тип данных, Структура 1, [Параметры])* – выбор из структуры элементов с учетом заданного набора параметров.

Для каждой операции задается имя структуры, сохраняющей результат, а после ее завершения устанавливается *статус* выполнения (1 – операция выполнена успешно; -1 – операция не выполнялась; -3 – ошибка сохранения результата и др.), который может быть использован как в отладочных целях, так и в операторе условного перехода. Тип структуры не влияет на выполнение операции на уровне пользователя, отличается только внутренняя реализация операций для различных структур.

Операции управления включают в себя:

1. *Копирование (Тип данных, Структура 1, [Параметры], Новая структура)* – копирование элементов одной структуры в другую с учетом заданного набора параметров.

2. *Условный переход (Статус, Знак, Значение, Номер операции для перехода)* – переход на указанную операцию в зависимости от статуса предыдущей.

3. *Комментарий* – строка для введения поясняющей информации, не влияющей на ход анализа.

Язык сценариев поддерживает сохранение сценариев и шаблонов, созданных на их основе. В качестве средства аналитической обработки создан гибкий инструмент, позволяющий на основе структур, полученных в результате обработки текста, строить различные ал-

горитмы анализа и изменять их при необходимости.

ИСПОЛЬЗОВАНИЕ ЯЗЫКА СЦЕНАРИЕВ ДЛЯ РЕШЕНИЯ ПРАКТИЧЕСКИХ ЗАДАЧ

Использование языка сценариев позволяет реализовать алгоритмы решения разнообразных задач, не создавая отдельные программные продукты, что демонстрируется на примерах решения распространенных на практике задач: составления словарей писателей и классификации текстов.

Например, для построения словарей писателей было отобрано и проанализировано около 200 текстов произведений разных писателей, по каждому из которых средствами подсистемы базовой обработки был построен словник, содержащий слово в начальной форме, его часть речи, абсолютную и относительную частоту использования в тексте.

Помимо построения словарей по большим объемам текстов, использование языка сценариев дает возможность сократить время анализа текстов, объемы которых не позволяют получить словник сразу всего текста или делают

Таблица 1

Результаты построения словарей писателей

№	Писатель	Размер фай-лов, Мб	Общий размер словника	Кол-во значимых частей речи	Время построения (мин:с)	
					Словников	
					Объединения	
					Всего	Словника
1	А. Азимов	3,27	20167	19970	01:17	
					00:49	
					02:06	2:19
2	И. Гончаров	4.9	22159	21952	01:26	
					00:32	
					02:58	04:05
3	А. Пушкин	1.53	12695	12520	00:42	
					00:06	
					00:48	00:57
4	Д. Донцова	5.19	26360	26138	02:38	
					00:47	
					03:25	04:48
5	Л. Толстой	7.58	24099	23833	03:30	
					01:22	
					04:52	06:03
6.	С. Лукьяненко	5.0	24821	24608	02:40	
					00:37	
					03:17	04:38
7.	М. Лермонтов	0.298	6018	5886	00:07	
					00:02	
					00:09	00:07
8.	Ф. Незнанский	2.4	21618	21413	01:13	
					00:20	
					01:23	2:03
9.	В. Пелевин	1.26	15836	15713	00:29	
					00:36	
					1:05	00:53
10.	Д. Черкасов	1.04	17088	16908	00:35	
					00:11	
					00:46	00:45

это затруднительным из-за больших затрат памяти и времени. Например, для текстов 30 произведений Л. Н. Толстого время построения словника сразу по всем текстам заняло 6 мин. 3 сек., тогда как использование сценария объединения словарей текстов позволило получить словарь Л. Н. Толстого за 4 мин. 52 сек. Обработка всех текстов требует большого количества ресурсов компьютера, что в данном случае и приводит к большему времени обработки. Использование сценариев не только дает возможность обрабатывать большие объемы текстовой информации с меньшими затратами памяти, но и в некоторых случаях сокращает время обработки (табл. 1). В случае независимости обрабатываемых структур в операциях сценария существует возможность их параллельного выполнения.

Для больших объемов текстовой информации актуальной является задача классификации текстов и автоматического рубрицирования. Язык сценариев с одной стороны позволяет решать непосредственно задачу автоматизации классифицирования текстов, с другой – задачу построения наборов ключевых слов для различных областей.

Для демонстрации данной возможности в качестве исходных данных были выбраны статьи по нескольким тематикам: анализ текста, биология, базы данных, компьютерная графика, теория и методика фигурного катания. Специально была выбраны часть тем из одной области науки (компьютерной) и часть из разных областей иллюстрации разницы распознавания в таких случаях.

По части текстов были построены словники, содержащие наборы ключевых слов для заданной области, остальные классифицировались на основе ключевых слов, построенных из ранее проанализированных текстов.

После построения словника по анализируемому тексту, средствами языка сценариев производится выбор ключевых слов этого текста. Для этого:

• из общего словника выбираются имена существительные;

• из полученного списка имен существительных с соответствующими им значениями частот выбираются слова с наибольшими значениями относительных частот, пороговое значение выбирается экспериментальным путем.

Сценарий, результатом выполнения которого является список ключевых слов текста, представлен на рис. 1.

В приведенном примере в качестве пороговых значений использовались 1 % и 0.5 %. Для других текстов это значение может варьироваться, в первую очередь в зависимости от объема текста.

После применения данного сценария к разным текстам из нескольких областей знаний, путем использования операций объединения возможно получение набора ключевых слов, характеризующих определенную область.

На основе построенных наборов ключевых слов, характерных для различных областей, может быть написан сценарий, определяющий к какой из имеющихся в системе областей может быть отнесен исследуемый текст. Для этого необходимо:

• найти набор ключевых слов исследуемого текста (рис. 1);

• найти отношение полученного набора ключевых слов к наборам ключевых слова каждой области в системе (рис. 2);

• выбрать область с наибольшим значением отношения.

В ходе эксперимента были проанализированы 8 текстов из 5 областей с использованием коэффициентов 0.01 (1 %) и 0.005 (0.5 %).

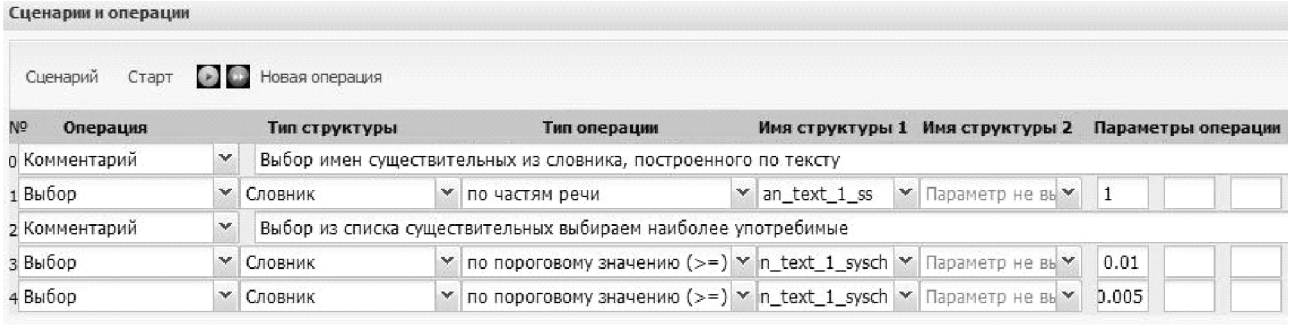
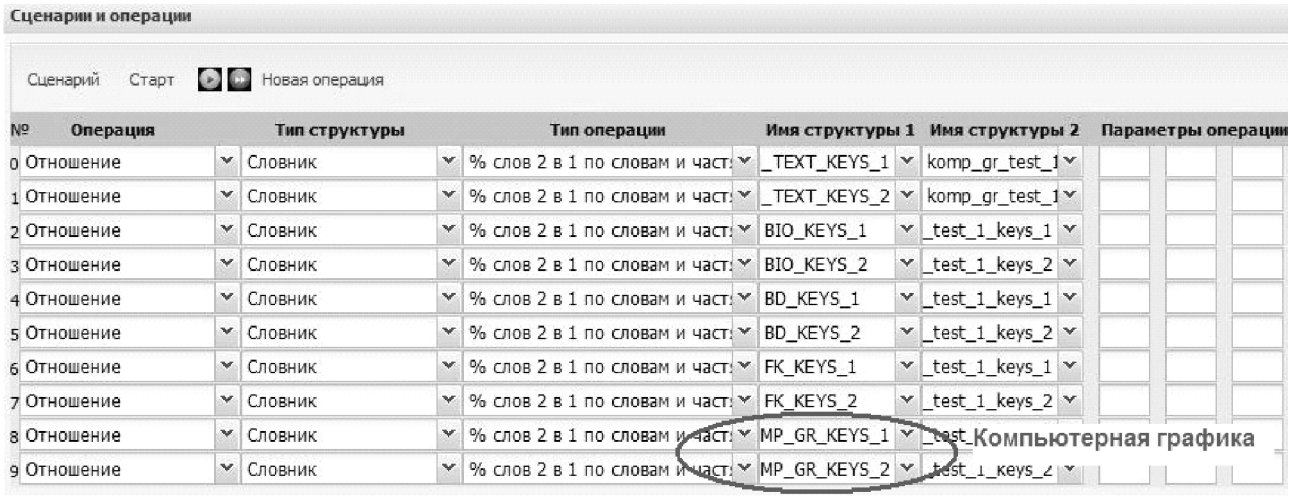


Рис. 1. Сценарий для построения списка ключевых слов текста


Сценарии и операции

Сценарий Старт Новая операция

№	Операция	Тип структуры	Тип операции	Имя структуры 1	Имя структуры 2	Параметры операции
0	Отношение	Словник	% слов 2 в 1 по словам и част	_TEXT_KEYS_1	komp_gr_test_1	
1	Отношение	Словник	% слов 2 в 1 по словам и част	_TEXT_KEYS_2	komp_gr_test_1	
2	Отношение	Словник	% слов 2 в 1 по словам и част	BIO_KEYS_1	_test_1_keys_1	
3	Отношение	Словник	% слов 2 в 1 по словам и част	BIO_KEYS_2	_test_1_keys_2	
4	Отношение	Словник	% слов 2 в 1 по словам и част	BD_KEYS_1	_test_1_keys_1	
5	Отношение	Словник	% слов 2 в 1 по словам и част	BD_KEYS_2	_test_1_keys_2	
6	Отношение	Словник	% слов 2 в 1 по словам и част	FK_KEYS_1	_test_1_keys_1	
7	Отношение	Словник	% слов 2 в 1 по словам и част	FK_KEYS_2	_test_1_keys_2	
8	Отношение	Словник	% слов 2 в 1 по словам и част	MP_GR_KEYS_1	_test_1_keys_1	
9	Отношение	Словник	% слов 2 в 1 по словам и част	MP_GR_KEYS_2	_test_1_keys_2	

Компьютерная графика

Рис. 2. Сценарий для нахождение отношений наборов ключевых слов

В результате в 6 случаях текст был классифицирован верно, в двух других отнесен к другой области: вместо “Анализа текста” и “Баз данных” текст был отнесен к области “Компьютерная графика”, при этом значения отношения для правильной области так же имело значение, большее чем для всех остальных областей. Это позволило сделать вывод, что для распознавания близких областей необходимо корректировать коэффициенты. Второй текст из области “Анализа текста”, три текста по “Базам данных” и тексты по “Компьютерной графике” и “Фигурному катанию” были правильно отнесены к своей области. Эксперимент проводился на базе 28 текстов. Увеличение количества проанализированных текстов для составления набора ключевых слов областей даст возможность оценить точность распознавания и на основе сделанной оценки усложнить алгоритм классификации.

ЗАКЛЮЧЕНИЕ

В статье предложен разработанный в составе открытой системы автоматизированного анализа текста язык сценариев, который дает возможность решения следующих классов задач:

1. *Лингвистические задачи:* составление словарей писателей, определение авторства, определение особенностей стиля писателя и т. д.

2. *Задачи по систематизации текстов:* автоматическая классификация, аннотирование, реферирование, в том числе выделение ключевых слов предметной области.

3. *Поисковые задачи:* поиск по информации, извлеченной из текстов.

4. *Задачи сравнения текстов:* определение плагиата, использования одного текста в другом и т. д.

5. *Задачи определения характеристик текстов:* определение статистических, лингвистических и интегральных характеристик текстов и структур, извлекаемых из них.

Язык сценариев – гибкий инструмент, предназначенный для интегрального анализа текста, позволяющий на основе структур, полученных в результате обработки, строить различные модели текста и легко изменять их впоследствии. Имея существенное преимущество перед закрытыми системами, открытая система автоматизированного анализа текста дает возможность сопоставления результатов, обмена алгоритмами между пользователями, накопления алгоритмов анализа и обобщения результатов их выполнения.

СПИСОК ЛИТЕРАТУРЫ

1. *Белоногов Г. Г., Калинин Ю. П., Хорошилов А. А., Хорошилов Ал-сей А.* Компьютерная лингвистика и перспективные информационные технологии. НТИ СЕР. 2 . Информ. процессы и системы. 2004 . № 8.

2. *Ермаков А.Е.* Извлечение знаний из текста и их обработка: состояние и перспективы. - Информационные технологии. – 2009. – N 7. – С. 50–55.

3. *Попов Э.В.* Общение с ЭВМ на естественном языке. Изд. 2-е, стереотипное. М.: Эдиториал УРСС, 2004. (Науки об искусственном).

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

5. *Балакирев Н.Е., Добрышина Е. В.* Язык сценариев для анализа информации, извлекаемой из текстов на естественном языке. – Тезисы докладов 9-ой Международной научно-методической конференции «Информатика: проблемы, методология, технологии». Воронеж, 2009 г. С. 51–52.

Балакирев Николай Евгеньевич – к. т. н., профессор кафедры проектирования вычислительных комплексов МАТИ – Российского государственного технологического университета имени К. Э. Циолковского. E-mail: balakirev1949@yandex.ru

Полицына Екатерина Валерьевна – к. т. н., ассистент кафедры проектирования вычислительных комплексов МАТИ – Российского государственного технологического университета имени К. Э. Циолковского. E-mail: kathrin.beaver@mail.ru

Balakirev Nikolay E. – candidate of technical sciences, professor, department of «Design of computing systems», Russian State Technological University named after K. E. Tsiolkovskiy. E-mail: balakirev1949@yandex.ru

Politsyna Ekaterina V. – candidate of technical sciences, assistant, department of «Design of computing systems », Russian State Technological University named after K. E. Tsiolkovskiy. E-mail: kathrin.beaver@mail.ru