УДК 004.912

ПРИМЕНЕНИЕ МАРКЕМНОГО АНАЛИЗА ДЛЯ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ В КОМПЛЕКСЕ ИНСТРУМЕНТОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

С. А. Полицын, Е. В. Полицына

Московский авиационный институт (национальный исследовательский университет)

Поступила в редакцию 30.03.2017 г.

Аннотация. В настоящее время объемы и динамика информации, которая подлежит обработке, делает особенно актуальной задачу автоматического выделения ключевых слов. В комплексе инструментов автоматизированного анализа текста реализован сервис выделения ключевых слов на основе построения частотного словаря словоформ. Каждая словоформа в тексте обладает двумя параметрами: частотой и длиной, поэтому на основе частотного распределения не всегда выделяются в качестве ключевых слова, отражающие суть текста. Маркемный анализ позволяет учитывать обе характеристики. Для его применения в инструментах выделения ключевых слов была расширена структура данных в открытой системе автоматизированной обработки текста и реализован новый сервис на портале «Автоматизированный анализ текста». Применение маркемного анализа в сервисе выделения ключевых слов и внедрение его поддержки в открытую систему автоматизированной обработки текстов дает возможность создавать сценарии для более точного решения различных практических задач и всестороннего исследования художественных текстов.
Ключевые слова: инструменты автоматизированного анализа текстов, выделение ключевых слов, применение маркемного анализа, сервис выделения ключевых слов.
Annotation. In the present time, the volumes and dynamics of the information flow to process make the automated key word extraction a very urgent task. There is the implementation of the keyword extraction service in the complex of automated text analysis tools based on the word frequency vocabulary of a text. Each word form in the text has two characteristics: its frequency and length, so the results of keyword extraction could be sometimes not very accurate basing on frequencies only. Markem analysis allows taking into account both characteristics. A new wordlist structure was introduced for the application of the keyword extraction service in the complex of automated text analysis tools, this service was also deployed to the portal "Automated Text Analysis". The use of markem analysis in the keyword extraction service and markem support in the open automated text processing system allows both creating text analysis scenarios for more accurate solving of practical tasks and comprehensive research of fiction literature.
Keywords: automated text analysis tools, keyword extraction, markem analysis, keyword extraction service.

ВВЕДЕНИЕ

Текст является сложным и многоаспектным объектом анализа. Задача компьютерного «понимания» текста является центральной в области автоматизированного анализа текстов на естественном языке. Эта задача состоит из множества составляющих, главной из которых можно назвать задачу выделения ключевых слов и понятий.

В настоящее время объемы и динамика информации, которая подлежит обработке, делает особенно актуальной задачу автоматического выделения ключевых слов. Это связано с применением интеллектуального анализа текстовых данных в широком спектре задач:

- информационный поиск;
- построение онтологий и др. семантических представлений;
- классификация и кластеризация;
- индексирование;

• аннотирование и реферирование;
• SEO и др.

В общем случае ключевыми могут быть слова, словосочетания, предложения. Под **ключевым** понимается слово в тексте, способное вместе с другими ключевыми словами передать тему текста. Ключевые слова являются носителями наиболее существенной информации в тексте. В настоящее время существуют статистические, лингвистические и гибридные (комбинированные) методы выделения ключевых слов.

Статистические методы основываются на численных данных встречаемости слова: частоте встречаемости слова в тексте, TF-IDF, частоте встречаемости сочетаний слов и др. Статистические методы далеко не всегда обеспечивают удовлетворительное качество результатов. К их преимуществам относятся универсальность, возможность быстрого использования для новых предметных областей; т. к. многие методы не требуют дополнительных данных, кроме исходного текста.

Лингвистические методы основываются на использовании онтологий и семантических данных о слове, лексико-синтаксических шаблонах, маркемном анализе и др. Лингвистические методы отличаются более высоким качеством получаемых результатов, но требуют наличия семантического представления, словарей и т.д., построение которых является крайне трудоемкими задачами, нуждающимися в автоматизации.

Гибридные (комбинированные) методы основываются на дополнении статистических одним или несколькими лингвистическими алгоритмами, например, дополнение расчетов статистики встречаемости слов использованием морфологических шаблонов или словарей синонимов и др.

В комплексе инструментов автоматизированного анализа текстов [1] используется статистический метод выделения ключевых слов, основанный на построении частотных словарей словоформ и их фильтрации.

ИНСТРУМЕНТЫ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ

В основу комплекса инструментов автоматизированного анализа текстов [1] положена *открытая система автоматизированной обработки текстов* на русском языке [2], которая включает в себя набор инструментов обработки текстов, накопления полученной информации и ее последующего анализа. Система является клиент-серверным многопользовательским веб-приложением, имеющим графический интерфейс пользователя для доступа к инструментам базовой и аналитической обработки и программный интерфейс (API) для обеспечения возможности использования инструментов системы в других приложениях.

Открытая система автоматизированной обработки текстов предоставляет 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 выделения ключевых слов, классификации и получения статистических характеристик текстов, представленных на *портале «Автоматизированный анализ текста»* (http://textanalysis.ru).

В открытой системе автоматизированного анализа текстов обработка производится на двух уровнях: базовой и аналитической обработки. **Структуры данных**, извлекаемые из текстов на этапе базовой обработки являются исходными данными для аналитической, на этапе которой могут быть реализованы различные алгоритмы анализа средствами языка сценариев. **Словники**, построенные по текстам на уровне базовой обработки, имеют следующую структуру: *слово в начальной форме; часть речи; абсолютная частота; относительная частота*.

После построения словника по анализируемому тексту, средствами **языка сценариев** [3] производится выбор **ключевых слов** этого текста: из общего словника выбираются имена существительные, затем из полученного списка имен существительных с соответству-

ющими им значениями частот выбираются слова с наибольшими значениями относительных частот, пороговое значение выбирается экспериментальным путем. Ключевыми словами при этом считается вершина списка (10–50 штук).

Алгоритм реализован средствами языка сценариев и используется в качестве шаблона как часть других сценариев. На основе его использования разработан **сервис получения ключевых слов**, внедренный на портале «Автоматизированный анализ текста». Результаты получения ключевых слов по тексту романа Е. И. Замятина «Мы» представлены на рис. 1. В целом, алгоритмы, базирующиеся на использовании частотных словарей, хотя и позволяют решать практические задачи выделения ключевых слов, классификации, кластеризации текстов и т. д., но все же даже на больших наборах данных далеко не всегда показывают хорошие результаты. Это связано с тем, что «частота слова в тексте – не простая величина, а *равнодействующая* двух закономерностей: объективной – языковой и субъективной – текстовой» [4], поэтому на основе частотного распределения не всегда выделяются в качестве ключевых слова, отражающие суть текста.

Каждая словоформа в тексте обладает двумя параметрами: частотой и длиной. «Частота словоформы является сложным показателем, а длина словоформы – простым, то субъективный фактор может быть получен простым вычитанием объективного фактора (веса словоформы по длине) из субъективно-объективного (веса словоформы по частоте). Полученная величина – Индекс Тематической Маркированности словоформы (ИнТеМ) – и будет указывать на степень субъективной (текстовой) весомости данной словоформы для данного текста» [4]

$$\text{ИнТеМ} = Q_{\text{вес}} - F_{\text{вес}}, \qquad (1)$$

где $Q_{\text{вес}}$ – параметрический вес слова по его частоте, а $F_{\text{вес}}$ – параметрический вес слова по его длине.

Для реализации поддержки маркемного анализа [4, 5] в открытой системе автоматизированной обработки текстов был расширен список существующих структур данных, добавлены **словники с ИнТеМом**, имеющие следующую структуру: *слово; слово в началь-*

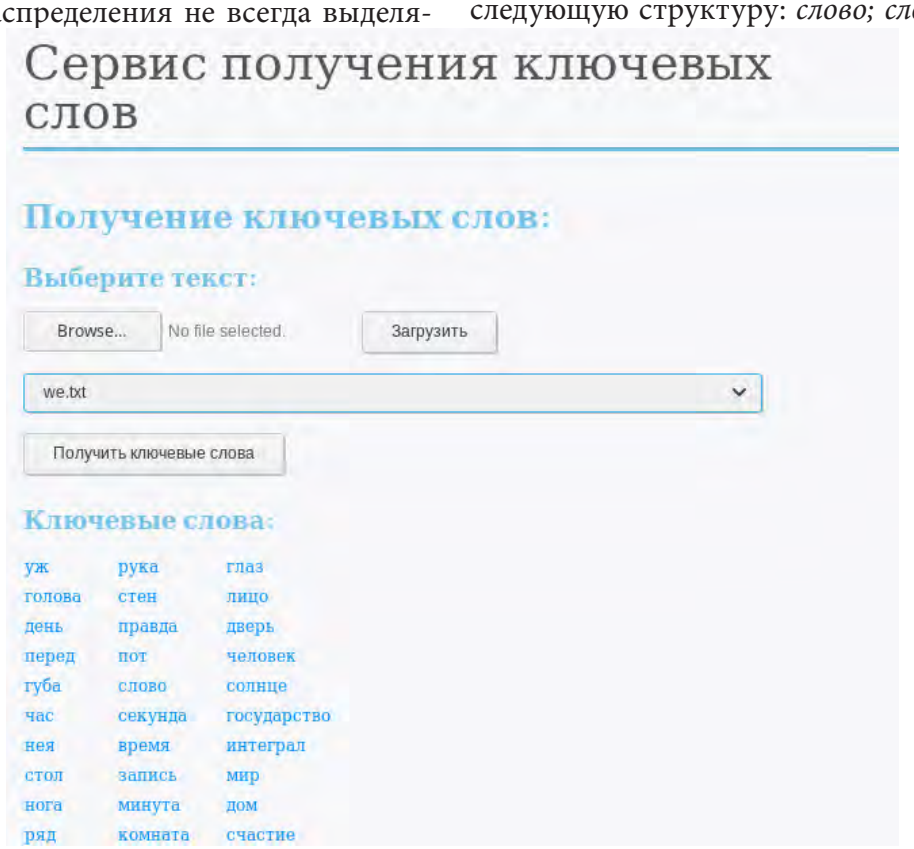

Рис. 1. Результаты использования сервиса выделения ключевых слов

ной форме; часть речи; длина слова; абсолютная частота; относительная частота; $Q_{вес}$*;* $F_{вес}$*; ИнТеМ.*

Исходными данными для вычисления ИнТеМа является словник, получаемый по исходному тексту на этапе лингвистической обработке в системе. В рамках общей концепции системы создание словников с ИнТеМом относится к разделу морфологии этапа статистической обработки (рис. 2).

Рис. 2. Добавление инструмента расчета словника с ИнТеМом в интерфейсе системы

Реализация алгоритма маркемного анализа в открытой системе автоматизированной обработки текста позволяет создать новый сервис выделения ключевых слов (рис. 3), учитывающий не только частоту употребления слов в тексте, но и их длину.

Результаты получения ключевых слов по тексту романа Е. И. Замятина «Мы» с помощью двух сервисов представлены в табл. 1. Пересечение полученных ключевых слов составляют только слова: *государство*, *бесконечность*, *конспект*, *страница*, *хранитель*, *интеграл*, – большая же часть полученных слов не совпадает, причем лучше отражают смысл романа ключевые слова, полученные с применением маркемного анализа.

ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ НА ОСНОВЕ МАРКЕМНОГО АНАЛИЗА

С помощью созданного сервиса выделения ключевых слов с ИнТеМом была произведена классификация текстов и проведено сравнение с результатами применения ранее реали-

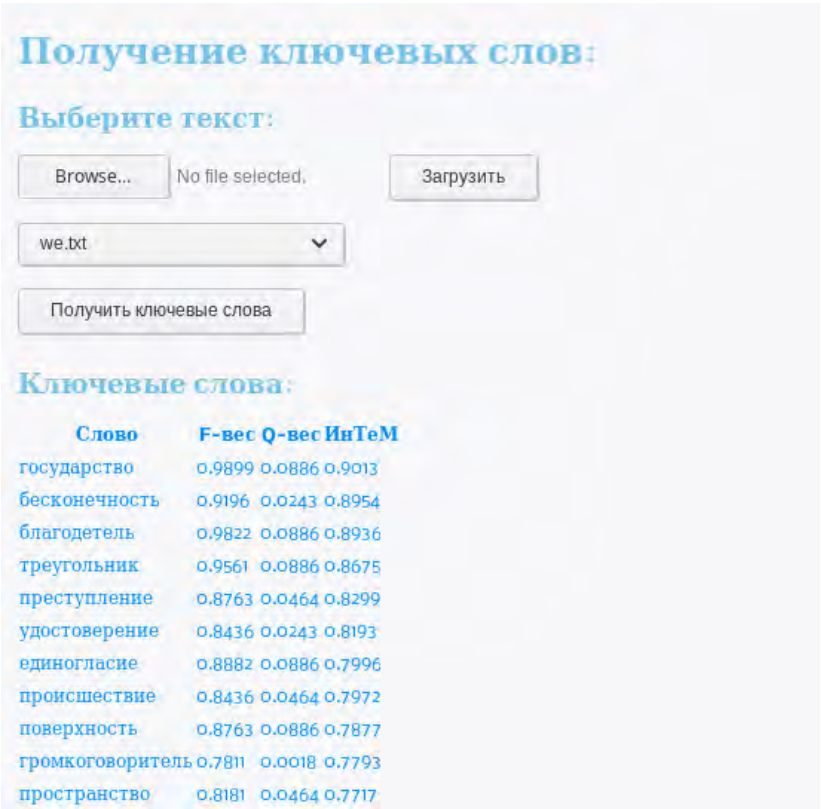

Рис. 3. Результаты использования сервиса выделения ключевых слов с поддержкой маркемного анализа

Таблица 1

Результаты получения ключевых слов по тексту романа Е. И. Замятина «Мы» с помощью двух сервисов

Сервис на основе частоты употребления слова			**Сервис на основе маркемного анализа**		
государство	строитель	настоящее	голова	**хранитель**	штора
бесконечность	инструмент	**конспект**	день	шаг	угол
благодетель	состояние	**страница**	перед	рука	тело
треугольник	вселенная	умывальник	губа	стен	машина
преступление	наслаждение	преступник	час	правда	**страница**
удостоверение	двигатель	мальчишка	стол	пот	сила
единогласие	чемоданчик	проспект	нога	слово	ночь
происшествие	христианин	скрижаль	ряд	секунда	глаз
поверхность	компрессия	бессмыслица	древние	время	лицо
громкоговоритель	полумесяц	любопытство	голос	запись	дверь
пространство	уравнение	впечатление	золотой	минута	человек
математика	маленькая	случайность	жизнь	комната	солнце
кают-компания	плоскость	растворение	конец	быль	**государство**
обязанность	**интеграл**	выключатель	ребенок	**благодетель**	**интеграл**
предрассудок	освобождение	операция	тень	**конспект**	мир
четырехугольник	революция	прогулка	право	круг	дом
хранитель	складочка		земля	улыбка	

Таблица 2

Результаты применения сервиса классификации текстов

	Анализ текста	**Базы данных**	**Фигурное катание**	**Компьютерная графика**	**Биология**
Анализ текста	**8**	5	1	**10**	0
Биология	3	0	1	1	**16**
Базы данных	3	**15**	3	**20**	0
Комп. графика	7	7	5	**35**	1
Фигурное катание	3	2	**14**	5	5

Таблица 3

Результаты классификации текстов с применением сервиса выделения ключевых слов с ИнТеМом

	Анализ текста	**Базы данных**	**Фигурное катание**	**Компьютерная графика**	**Биология**
Анализ текста	**20**	18	14	**22**	6
Биология	6	10	8	2	**22**
Базы данных	12	**18**	4	10	6
Комп. графика	16	20	6	**24**	8
Фигурное катание	16	16	**50**	18	14

Таблица 4

Общие ключевые слова по результатам применения двух методов

Тексты	Пересечение ключевых слов	Процент пересечения, %
Базы данных	разработка редактирование приложение сортировка управление структура обработка инструмент	16
Биология	наследование изменение результат количество скрещивание полисахарид расщепление соединение	16
Фигурное катание	выносливость квалификация исполнение композиция увеличение подготовка мастерство выступление скольжение координация воспитание	22
Компьютерная графика	устройство насыщенность зависимость разрешение результат изображение количество параметр значение диапазон смешение информация пространство представление	28
Астрономия	объект сближение область Юпитер посол спутник солнце	29
Анализ текста	рубрицирование существительное предложение результат рубрикация сочетание программа выражение содержание произведение исследование рубрикатор обработка словосочетание информация выделение	32
Собрание сочинений М. В. Булкагова		0
Собрание сочинений Л. Н. Толстого	главнокомандующий воображение	4
Л. Н. Толстой "Война и мир"	главнокомандующий	2
Е. И. Замятин "Мы"	государство хранитель страница благодетель интеграл конспект	12

зованного в системе сценария классификации текстов и сервиса на портале. Исследовались статьи по нескольким тематикам: анализ текста, биология, базы данных, компьютерная графика, теория и методика фигурного катания. В табл. 2 приведены ранее полученные результаты [5], в табл. 3 – результаты, полученные с применением маркемного анализа.

Результаты классификации, полученные с применением маркемного анализа, больше отражают истинную картину, при этом процент пересечения ключевых слов, полученных двумя методами для научных текстов в среднем составляет 24 %, для художественных текстов около 4%. В результатах, полученных каждым методом проявляются разные особенности естественно-языковых текстов. В табл. 4 приведены результаты пересечения ключевых слов, полученных двумя методами. В списках общих ключевых слов содержится в том числе общенаучная лексика, что приводит к близости полученных значений в разных областях, например, 20 % для области «Анализ текста» и 22 % для области

«Компьютерная графика» для исследуемого текста из области «Анализ текста».

Например, пересечением ключевых слов предметной области «Анализ текста» с ключевыми словами области «Компьютерная графика» являются: **технология**, **необходимость**, компонент, изображение, содержание, **возможность**, **информация**, **пользователь**, **использование**, **представление**.

Поскольку «лексика любого текста может быть разделена на две части: одна из них связана с темой текста, другая – никак не указывает на тему текста и может встретиться в тексте любой другой тематики» [4], то для решения этой проблемы необходимо применять дополнительные фильтры [4, 5], как при формировании списков ключевых слов предметных областей, так и при выделении ключевых слов исследуемых текстов.

РАЗВИТИЕ ИНСТРУМЕНТОВ КОМПЛЕКСА АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТОВ

Анализ результатов классификации текстов двумя методами показал целесообразность внедрения полной поддержки маркемного анализа в открытую систему автоматизированной обработки текста и развитие инструментов комплекса на его основе. Помимо реализации алгоритма вычисления ИнТеМа и создания сервиса выделения ключевых слов для возможности его последующего использования в алгоритмах классификации, кластеризации в языке сценариев необходимо расширить список операций, реализовав для словников с ИнТеМом следующие: *объединение* (по словам, по словам и частям речи, с пересчетом ИнТеМа); *пересечение* (по словам, по словам и частям речи, с заданной близостью значений ИнТеМа); *разность* (по словам, по словам и частям речи, с заданной близостью значений ИнТеМа); отношение (% слов 2 в 1 по словам, % слов 2, отличных от 1 по словам, % слов 2 в 1 по словам и частям речи, % слов 2, отличных от 1 по словам и частям речи); *объединение с отсечением* с учетом длин слов (>=, <=, диапазон), с учетом значений ИнТеМа (<=, >=, диапазон), по количеству слов); *удаление* (по частям речи, по значению ИнТеМа (<=, >=, диапазон), по количеству слов, по длине слова (<=, >=, диапазон)); *выбор* (по частям речи, по значению ИнТеМа (<=, >=, диапазон), по количеству слов, по длине слова (<=, >=, диапазон)). Внедрение нового вида структур и операций над ними позволит реализовывать сценарии разных алгоритмов выделения ключевых слов, классификации текстов, исследования текстов и т. д.

Таким образом, применение маркемного анализа для выделения ключевых слов дает возможность развития инструментов комплекса в следующих **направлениях**:

1. Расширение алгоритма выделения ключевых слов: создание фильтров общенаучной лексики, предметных областей и т. д., учет синонимов, комбинирование с другими лингвистическими методами.

2. Реализация полной поддержки операций над словниками с ИнТеМом в языке сценариев.

3. Замена алгоритма выделения ключевых слов в сервисе классификации.

4. Реализация алгоритма выделения ключевых слов внутри предметных областей с применением автоматической классификации.

5. Реализация алгоритма выделения ключевых словосочетаний и предложений.

ЗАКЛЮЧЕНИЕ

Применение маркемного анализа для выделения ключевых слов в комплексе инструментов автоматизированного анализа текстов на русском языке, внедрение его полной поддержки в открытую системы автоматизированной обработки текста позволит расширить возможности его применения, создавать сценарии анализа для более точного решения практических задач, всестороннего исследования художественных текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Балакирев Н. Е., Полицына Е. В.* Подход к созданию комплекса инструментов автоматизированного анализа текстов на русском

языке // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2016. – № 2. –С. 98–105.

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

3. *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.

4. *Кретов А. А.* Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2007. – № 1. – С. 81–90.

5. *Кретов А. А.* Понятие маркемы: методика выявления и практика использования / А. А. Кретов // Универсалии русской литературы. 2. – Воронеж: Научная книга, 2010. – С. 138–153.

6. Б*алакирев Н. Е., Полицына Е. В.* Реализация адаптивно-динамической модели преобразования информации средствами языка сценариев на примере задачи классификации текстов // Материалы XI Международной научно-методической конференции "Информатика: проблемы, методология, технологии". Т. 1. – Воронеж. – 2011. – С. 73–77.

Полицын Сергей Александрович – ст. преподаватель, институт № 4, кафедра «Проектирование вычислительных комплексов», Московский авиационный институт (Национальный исследовательский университет).
Тел.: 8-499-141-94-82
E-mail: pul_forever@mail.ru

Полицына Екатерина Валерьевна – канд. техн. наук, доцент, институт № 4, кафедра «Проектирование вычислительных комплексов», Московский авиационный институт (Национальный исследовательский университет).
Тел.: 8-499-141-94-82
E-mail: kathrin.beaver@mail.ru

Politsyn Sergey A. – senior lecturer, department of «Design of Computing Systems», Moscow Aviation Institute (National Research University).
Tel.: 8-499-141-94-82
E-mail: pul_forever@mail.ru

Politsyna Ekaterina V. – candidate of technical sciences, associate professor, department of «Design of Computing Systems», Moscow Aviation Institute (National Research University).
Tel.: 8-499-141-94-82
E-mail: kathrin.beaver@mail.ru