УДК 004.912, 51-74

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РУССКОЯЗЫЧНОГО ТЕКСТОВОГО ДОКУМЕНТА ДЛЯ РЕШЕНИЯ ЗАДАЧИ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ ТЕКСТА

А. С. Петров, Т. Э. Шульга

Саратовский государственный технический университет им. Ю. А. Гагарина

Поступила в редакцию 28.08.2017 г.

Аннотация. Предложена математическая модель русскоязычного текстового документа, которая может быть использована для извлечения ключевых слов и словосочетаний из текстовых корпусов с применением лингвистических фильтров, списка стоп-слов и статистических метрик. Приведено описание функционала программного обеспечения, реализующего известные методы извлечения терминов C-value и k-factor. Проведен сравнительный анализ реализованных в описываемом ПО статистических методов на русскоязычных текстах исторической предметной области.
Ключевые слова: автоматическое извлечение терминов, лингвистические фильтры, статистические методы, математическая модель текстового документа.
Annotation. The model of a russian-language text document, which can be used in the task of term extraction from corpora, is offered. The software implementation of the C-value and k-factor methods is described. The results of the methods comparison using the russian-language texts of historical domain are presented.
Keywords: automatic term recognition, linguistic filters, statistical methods, the text document mathematical model.

ВВЕДЕНИЕ

Автоматическое извлечение ключевых слов и терминов из текста является важным этапом в решении ряда актуальных задач, связанных с обработкой текстов определенной предметной области, среди которых – информационный поиск, классификация, кластеризация документов, полуавтоматическое построение онтологий. В частности, при полуавтоматическом построении онтологий на основе текстовых коллекций необходимо выделить из текста основные понятия (термины) и связи между ними. Рассмотрим в качестве примера предложение, взятое из исторического военного документа: «При вклинении противника в передний край обороны минометы своим огнем поражают противника в занятых им районах» [1]. Терминами предметной области в данном предложении являются «противник», «передний край обороны», «миномёт» и «огонь».

В последние годы проблеме автоматического извлечения терминов из текста посвящено большое количество научных работ. Так, например, работы А. О. Шелманова и др. посвящены исследованию метода извлечения терминов из текста, комбинирующего лексическую, морфологическую, синтаксическую и семантическую информацию, применительно к задаче анализа научных публикаций [2, 3]. В работе А. О. Шелманова и И. В. Смирнова описано применение методов машинного обучения, а также продукционных правил для выделения понятий медицинской предметной области [4]. Работа Р. Е. Суворова и И. В. Соченкова посвящена применению методов распознавания ключевых слов для решения задачи определения связанности научно-технических документов [5]. Также существует ряд работ, посвященных сравнительному обзору методов автоматизированного извлечения терминов для различных корпусов, составленных из текстов различных языков мира. Как отмечают Э. С. Клышинский и Н. А. Кочеткова [6], наиболее эффективные

для русскоязычных текстов методы были описаны и оценены в работах П. И. Браславского и др. [7] и Н. В. Лукашевича и др. [8]. В работе [7] была исследована возможность выделения многословных терминов методами извлечения терминов. В качестве текстового корпуса применялись тексты генетической предметной области. В работе [8] на текстах естественнонаучной предметной области исследованию подверглась возможность извлечения однословных терминов.

В ходе анализа работ по данной тематике было установлено, что исследования применимости методов автоматического извлечения ключевых слов и терминов на текстах исторической предметной области ранее не проводились. Кроме этого, в научной литературе не удалось найти описание формализованной постановки задачи извлечения терминов из текста.

Целью данной работы является проведение исследования статистических методов выделения понятий на текстах исторической предметной области. В данной работе приведена содержательная постановка задачи автоматического извлечения терминов из текста и предложена математическая модель русскоязычного текстового документа, которая может быть использована для рассматриваемой задачи. Кроме этого, дано описание функционала разработанного программного обеспечения, реализующего известные методы извлечения ключевых слов и словосочетаний C-value и k-factor, приводится сравнительный анализ их применимости к текстовым документам исторической предметной области.

ЗАДАЧА ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ ТЕКСТА

Задача автоматического извлечения терминов из текста заключается в получении из исходного текстового корпуса перечня слов и словосочетаний, являющихся терминами определенной предметной области. Текстовый корпус представляет собой сформированную по определенным правилам, структурированную коллекцию текстов. Как правило, они составляются из текстов определенной предметной области. Например, текстовый корпус может включать в себя военно-исторические документы периода 1941–1945 гг. Под термином принято понимать слово или словосочетание на естественном языке, описывающее понятие определенной предметной области [9]. Заметим, что в научных работах, посвященных данной тематике, указанное выше определение приводится только на содержательном уровне. Это в некоторой степени затрудняет его трактовку при разработке математических методов и соответствующего программного обеспечения. Поэтому для решения задачи автоматического извлечения терминов из текста представляется целесообразным ввести математическую модель текстового документа на русском языке, учитывающую ряд присущих ему особенностей.

Пусть D – текстовый документ определенной предметной области,

$$D = \{s_1, \ldots, s_i, \ldots, s_n\}, \quad (1)$$

где $s_i = \{w_{i1}, sep'_{i1}, \ldots, sep'_{ib}, w_{im}, sep_{i1}, \ldots sep_{ik}\}$ – предложение (2)

i - номер предложения, $i = \overline{1,n}$

n - количество предложений в текстовом документе D

$$w_{ij} = \{l_{ij1}, \ldots, l_{ija}\} \text{ – слово} \quad (3)$$

где j – номер слова в предложении s_i, $j = \overline{1,m}$

m – количество слов в предложении s_i

$l_{ija'}$ – буква в слове w_{ij}

a' – номер буквы в слове w_{ij}, $a' = \overline{1,a}$

a – количество букв в слове w_{ij}

$sep_{ik'}$ – разделитель предложения w_i, $sep_{ik'} \in Sep$

$sep'_{ib'}$ – разделитель слова в предложении w_i, $sep'_{ib'} \in Sep'$

Sep и Sep' – множества символов-разделителей. Элементами множества Sep являются символы конца предложения, в то время как членами множества Sep' являются символы-разделители слов в предложении.

$$Sep = \{'.', '!', '?'\} \quad (4)$$

$$Sep' = \{' ', ',', ';', '-', ':'\} \quad (5)$$

k' – номер разделителя конца предложения, $k' = \overline{1,k}$

k – количество символов конца предложения s_i

b' – номер разделителя слова, $b' = \overline{1,b}$

b – количество разделителей слова w_{ij} в предложении s_i.

Каждое слово обладает морфологической парадигмой Wf – системой словоформ, образующих одну лексему [10]. Лексема – это единица словаря языка, объединяющая разные формы одного слова (метод, метода, методу и т. д.), а также разные смысловые варианты слова, зависящие от контекста. Таким образом, Wf – это множество словоформ одного слова.

$$Wf = \{wf_1, \ldots, wf_e \mid DNF(wf_{e'}) = wf_{nf}\} \quad (6)$$

для $\forall wf_{e'}, DPos(wf_{e'}) \in Pos', Pos' \in Pos$

$$wf_{e'} = (w, case, n) \text{ – словоформа} \quad (7)$$

где Pos – множество частей речи

$$Pos = (\text{'}Существительное\text{'}, \text{'}Местоимение\text{'}, \text{'}Прилагательное\text{'}, \text{'}Предлог\text{'}, \text{'}Частица\text{'}, \text{'}Союз\text{'}, \text{'}Наречие\text{'}, \text{'}Числительное\text{'}, \text{'}Причастие\text{'}, \text{'}Деепричастие\text{'}, \text{'}Глагол\text{'}) \quad (8)$$

Pos' – множество именных и местоименных частей речи

$$Pos' = (\text{'}Существительное\text{'}, \text{'}Местоимение\text{'}, \text{'}Прилагательное\text{'}, \text{'}Причастие\text{'}, \text{'}Числительное\text{'}) \quad (9)$$

Словоформа $wf_{e'}$ состоит из трех элементов:

w – слово, $w \in W$,

$case$ – падеж, $case \in Case$,

n – число, $n \in N$

$$Case = (\text{'}именительный\text{'}, \text{'}родительный\text{'}, \text{'}дательный\text{'}, \text{'}винительный\text{'}, \text{'}творительный\text{'}, \text{'}предложный\text{'}) \quad (10)$$

$N = (\text{'}единственное\text{'}, \text{'}множественное\text{'})$

e – количество словоформ одной парадигмы

e' – номер словоформы, $e' = \overline{1,e}$

У каждого слова w имеется лемма wf_{nf}. Лемма – это исходная, базовая или нормальная форма слова, зафиксированная в словаре [11]. Для именных и местоименных частей речи нормальной является форма именительного падежа единственного числа. Лемма является одной из словоформ слова w. Таким образом,

$$wf_{nf} = (w, case', n'), \quad wf_{nf} \in Wf, \quad (11)$$

где $case' = \text{'}именительный\text{'}, n' = \text{'}единственное\text{'}$.

Перейдем к описанию указанных функций. Обозначим $DPos$ функцией определения части речи слова, DNF – функцией нормализации слова. Нормализация слова или лемматизация – это процесс приведения слова к лемме или начальной форме [11].

$$DPos : W \to Pos \quad (12)$$

$$DNF : Wf \to Wf_{nf}. \quad (13)$$

Перейдем к описанию понятия терминологического кандидата. Кандидат в термины – слово или словосочетание, удовлетворяющее заданным критериям и потенциально являющееся термином определенной предметной области. Пусть P – множество терминологических кандидатов. Элементами данного множества являются слова w и словосочетания p.

$$P = \{w_1, \ldots, w_o, p_1, \ldots, p_r\}, \quad (14)$$

где o – количество кандидатов-слов,

r – количество кандидатов-словосочетаний,

$$p_u = \{w_{u1}, \ldots w_{uc} \mid c = \overline{1,z}\}, \quad (15)$$

где u – номер словосочетания, $u = \overline{1,r}$

c – номер слова в словосочетании p_u,

z – максимальная длина словосочетания; эмпирически установлено, что $z \le 5$.

Большинство методов решения задачи автоматического извлечения терминов включают два этапа. На первом этапе производится извлечение кандидатов в термины из текстового корпуса. Опишем функцию DP, выполняющую извлечение перечня терминологических кандидатов P из текстового документа D.

$$DP : D \to P. \quad (16)$$

На втором этапе путем фильтрации и ранжирования списка, полученного на предыдущем этапе с помощью функции DP, формируется результирующий список терминов. На втором этапе применяются, в частности, статистические методы и методы машинного обучения. Суть статистического метода состоит в подсчете метрики M с помощью функции нахождения значения статистической метрики F_m.

M – статистическая метрика, $M \in R, M \geq 0$

$$F_m : P \to M. \quad (17)$$

Описав все необходимые понятия и функции, перейдем к формальной постановке задачи извлечения терминов из текста.

Дано:

$TD = \{d_1, \ldots, d_{d''}\}$ – текстовый корпус (18)

где d'' – количество текстовых документов в текстовом корпусе

Найти:

$T = \{(p,\ m)_{v'}\}$ – множество пар: терминологические кандидаты и их значения метрик (19)

где v' – номер кандидата в текстовом документе, $v' = 1, v$

v – количество терминов в текстовом документе.

Множество T проранжировано в порядке убывания значений m. Таким образом, чем ближе к началу в итоговом множестве T находится кандидат p, тем с большей долей вероятности он является термином заданной предметной области.

ИССЛЕДУЕМЫЕ МЕТОДЫ

Предложенная математическая модель может быть использована для анализа существующих методов выделения ключевых слов из текста, а также для разработки новых методов. В данной работе проведен сравнительный анализ двух известных методов выделения терминов произвольной структуры: C-value и k-factor [12, 13]. Эти методы были выбраны, так как по результатам исследования на корпусе научных статей генетической предметной области они дают лучшие результаты [7].

Для выполнения этапа извлечения терминологических кандидатов данные методы не используют словари, онтологии или какие-либо другие семантические ресурсы. Кратко рассмотрим каждый из них.

Метод C-value базируется на использовании такой статистической метрики, как частота встречаемости строки в тексте. По сравнению с ней метрика C-value учитывает длину и вложенность терминологического кандидата. Вложенные термины (nested terms) – это понятия, содержащиеся в исходном тексте как по-отдельности, так и в составе других понятий [12]. Метрика, используемая методом C-value, подсчитывается согласно следующей формуле.

$$C - value(a) = \begin{cases} \log_2 |a| \cdot f(a), \\ \quad a - не\ вложен \\ \log_2 |a| \left(f(a) - \dfrac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right), \\ \quad a - вложен \end{cases} \quad (20)$$

где a – терминологический кандидат,

$|a|$ – длина a, выраженная в количестве слов

$f(\cdot)$ – частота встречаемости кандидата,

T_a – множество извлеченных кандидатов, содержащих a

$P(T_a)$ – количество кандидатов в T_a

$\sum f(b)$ – сумма частот встречаемости кандидатов $b \in T_a$, содержащих a. То есть a является вложенным кандидатом по отношению к b.

Из вышеописанной формулы можно сделать вывод, что чем длиннее строка a, тем больше значение ее метрики. Это сделано для учета следующей закономерности. Более длинные строки встречаются в исходном тексте реже коротких. Следовательно, вероятность появления строки b в количестве f упоминаний меньше, чем вероятность появление строки a в количестве f раз, при условии, что $|a| < |b|$. По этой причине можно сделать вывод, что словосочетание b с большей вероятностью является термином по сравнению с a. Кроме этого, данный метод создан с предположением, заключающемся в том, что чем выше количество T_a – строк, содержащих a, тем больше степень независимости a.

Метод, описанный в работе [13], не имеет собственного названия. В исследовании Браславского и др. [7] указанный метод фигурирует под названием k-factor. Данный метод реализован в системе BootCaT, предназначенной для формирования текстового корпуса из Веба с применением поисковой системы.

В качестве входного параметра используется перечень исходных однословных терминов, упоминающийся в работе под названием seed terms. Создаются запросы к поисковой системе, содержащие данные исходные термины, и, таким образом, извлекаются их содержащие текстовые документы. Полученные текстовые документы обрабатываются алгоритмом, в результате чего на выходе получается новое множество однословных терминов, которое используется для следующей итерации работы алгоритма, начинающейся с запроса к поисковой системе. Итоговый текстовый корпус используется для извлечения многословных терминов, для чего применяется рассматриваемый статистический метод. Метод k-factor базируется на использовании статистической метрики частоты встречаемости строки в тексте и, также как и метод C-value, учитывает вложенность терминологических кандидатов. Метод функционирует следующим образом: если строка a является вложенной по отношению к b и $f_a > \left(\frac{1}{k}\right) \cdot f_b$, то из двух кандидатов в финальном перечне терминов метод оставляет b.

ОПИСАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Алгоритм, реализованный в разработанном программном обеспечении, базируется на алгоритме, описанном в работе [12]. Данный алгоритм выполняет следующую последовательность этапов:

1. Для каждого слова из исходного текстового корпуса выполняется морфологический разбор.

2. Применяется лингвистический фильтр, предназначенный для ограничения типов слов и словосочетаний [9]. В данной работе исследуется работа двух лингвистических фильтров: $Noun^+$ [12, 14] и $(Adj \mid Noun)^+ Noun$ [12, 15].

Ограничение первого из них, $Noun^+$, состоит в том, что в качестве допустимых кандидатов могут выступать существительные и словосочетания существительных. Для второго фильтра $(Adj \mid Noun)^+ Noun$ допустимыми являются именные группы, то есть словосочетания, в которых главным словом является существительное (в частности, прилагательное + существительное).

3. Полученный список кандидатов фильтруется с применением списка стоп-слов – перечня слов, появление которых не является ожидаемым в списке терминов указанной предметной области [9].

4. Подсчитывается статистическая метрика. Итоговое множество слов и словосочетаний ранжируются в порядке убывания значений метрики.

Программное обеспечение создано с применением языка программирования Python. Опишем некоторые особенности реализации.

Исходный текстовый корпус может быть считан как из единственного файла, так и из множества файлов. Допустимыми типами являются текстовые файлы и pdf-документы. Для выполнения этапа № 1, заключающегося в морфологическом разборе каждого слова, применяется библиотека pymorphy2 [16]. Данная библиотека позволяет не только установить часть речи и форму слова, но также и преобразовать слово в любую форму, в том числе нормальную. Исходный текст делится на предложения (s_i), каждое из которых преобразуется в массив слов с тегами (w_{ij}) и разделителями слов в предложении $(sep'_{ib'})$. Результирующий перечень предложений обрабатывается лингвистическим фильтром. Таким образом, формируется список терминологических кандидатов (P), который потом проходит процедуру фильтрации с помощью стоп-списка. Для каждого кандидата подсчитывается частота встречаемости, а также определяется иерархия вложенности.

При запуске программы имеется возможность выбора входного файла, в котором хранится исходный текстовый корпус, а также той его части, которая будет впоследствии обработана программой. Также имеется возможность выбора типа лингвистического фильтра, а также методов подсчета статистической метрики для каждого кандидата. По мере выполнения программы формируется журнал событий, в котором отображается вся информация о ходе текущего выполнения

программы. По окончанию работы программы формируется ряд текстовых файлов, в каждом из которых записываются результаты выполнения одного выбранного метода.

МЕТОДИКА ЭКСПЕРИМЕНТА

Для проведения исследования был сформирован корпус из текстов исторической предметной области, а именно документы 1941–1945 гг., опубликованные в рамках двенадцати выпусков «Сборника боевых документов Великой Отечественной Войны», выпущенных в период с 1947 по 1950 гг. [1]. Суммарно текстовый корпус, использовавшийся в рамках данной работы, содержит 308 документов, 32836 предложений и 443764 слова. Корпус обрабатывался как единый текст, без разбивки на отдельные документы.

Для оценки результирующего списка извлеченных слов и словосочетаний в данной работе использовалась несколько видоизмененная методика оценки, описанная в работе [7]. В указанном исследовании была применена комбинация экспертной и формальной оценок. Для данной работы была проведена только формальная оценка результата.

Методика заключается в оценке трех списков: «короткого», «среднего» и «длинного». «Длинный список» – это полный результирующий список слов и словосочетаний, извлеченный с помощью разработанного программного обеспечения. «Средний список» состоит из строк, значение метрики которых больше единицы. Для метода k-factor элементами «короткого списка» являются 100 случайно выбранных с учетом длин строк из полного перечня результатов. Для метода C-value «короткий список» составляется из 100 верхних элементов отсортированного по убыванию результирующего перечня.

Для проведения формальной оценки используется словарь военных терминов [17]. Формальная оценка проводится на основе как четкого, так и нечеткого сравнения. Четкое сравнение проводится по трем параметрам: точное совпадение извлеченного методом термина a со словарным термином b, вложенность термина a по отношению к b (включение), вложенность термина b по отношению к a (вхождение). Нечеткое сравнение проводится по критерию близости двух строк. Данных критерий определяется как отношение числа совпавших слов к общему количеству уникальных слов двух строк:

$$sim(a,b) = \frac{|a \cap b|}{|a \cup b|}. \quad (21)$$

При нечетком сравнении проводится подсчет количества выделенных терминов $a \in T_a$, для которых

$$\exists\, b,\ sim(a,b) \geq 0{,}5. \quad (22).$$

В данной работе исследованию подвергались два метода извлечения терминов из текста C-value и k-factor. Кроме этого изучалось влияние типа лингвистического фильтра на итоговый результирующий список на примере двух лингвистических фильтров: $Noun^+$ и $(Adj \mid Noun)^+ Noun$.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Результаты формальной оценки короткого, среднего и длинного списка отображены в табл. 1, 2 и 3, соответственно.

Из результатов оценки полного списка можно сделать вывод, что при применении лингвистического фильтра $(Adj \mid Noun)^+ Noun$ длина извлеченного результирующего списка в среднем в 2,5 раза больше по сравнению с перечнем, полученным с помощью фильтра $Noun^+$. Помимо большего размера перечня слов и словосочетаний, извлеченных с применением фильтра $(Adj \mid Noun)^+ Noun$, данный фильтр позволяет извлечь большее количество терминов, согласно нечеткой и четким оценкам.

Сравнивая результаты оценки точного совпадения элементов полного и среднего списков с терминами словаря, можно увидеть, что 19 % терминов, полученных методом C-value, оказались за пределами среднего списка. Отношение количества точно совпавших терминов, извлеченных методом k-factor, по полному и среднему списку равно 2,3. Анализируя результаты нечеткой формальной оценки следует отметить, что для метода C-value про-

Таблица 1

Результаты формальной оценки короткого списка

Оценка \ Фильтр	C-value		k-factor	
	$Noun^+$	$(Adj \mid Noun)^+ Noun$	$Noun^+$	$(Adj \mid Noun)^+ Noun$
точно	7	27	1	0
включение	89	78	63	60
вхождение	7	23	3	1
нечеткая	43	68	4	5

Таблица 2

Результаты формальной оценки среднего списка

Оценка \ Фильтр	C-value		k-factor	
	$Noun^+$	$(Adj \mid Noun)^+ Noun$	$Noun^+$	$(Adj \mid Noun)^+ Noun$
размер списка	11355	37549	4471	9174
точно	100	341	79	95
включение	8782	27944	2425	5698
вхождение	119	250	100	132
нечеткая	1165	2878	668	1081

Таблица 3

Результаты формальной оценки длинного списка

Оценка \ Фильтр	C-value		k-factor	
	$Noun^+$	$(Adj \mid Noun)^+ Noun$	$Noun^+$	$(Adj \mid Noun)^+ Noun$
размер списка	18635	47308	26574	71989
точно	123	410	183	266
включение	12614	32117	15648	46313
вхождение	184	328	254	338
нечеткая	1964	4123	1578	2627

центное соотношение количества терминов в среднем списке к их количеству в полном списке составляет: 59,3 %, с применением фильтра $Noun^+$, и 69,8 %, с применением фильтра $(Adj \mid Noun)^+ Noun$. Для метода k-factor данное соотношение составляет 42,3 % и 41,1 %, соответственно. Из этого можно сделать вывод, что результирующие термины, полученные методом C-value, располагаются ближе к началу списка, по сравнению с выходным перечнем, извлеченным методом k-factor.

Сравнивая результаты данного исследования с исследованием [7], проведенным на русскоязычных текстах генетической предметной области, можно сделать вывод о сопоставимости результатов работы методов C-value и k-factor применительно к русскоязычным текстовым документам исторической предметной области. Однако, стоит отметить, что сравнивая результаты оценки «длинного списка», в работе [7] количество извлеченных методом k-factor терминов больше списка понятий, полученных методом C-value. В данной работе длина перечня выделенных методом k-factor терминов меньше по сравнению со списком, полученным при использовании метода C-value.

ЗАКЛЮЧЕНИЕ

В ходе исследования статистических методов выделения терминов была установлена их применимость на русскоязычных текстах исторической предметной области. На основе анализа результатов можно сделать вывод, что, несмотря на то, что метод k-factor позволяет извлечь большее количество слов и словосочетаний из исходного текста по сравнению с методом C-value, формальная оценка выделенных терминологических кандидатов показывает, что метод C-value позволяет получить большее количество терминов. При этом следует отметить характер их распределения в списке: метод C-value располагает термины предметной области ближе к началу списка по сравнению с методом k-factor. При сравнении лингвистических фильтров $Noun^+$ и $(Adj \mid Noun)^+ Noun$ было выявлено, что применение второго рассмотренного фильтра является более целесообразным в силу большего количества извлекаемых терминов. Данные результаты могут быть использованы исследователями, занимающимися проблемами анализа русскоязычных текстовых документов. В дальнейшем исследования методов решения задачи извлечения терминов из текста на русскоязычных текстах исторической предметной области будут продолжены для методов GlossEx [18] и TermExtractor [19]. Авторы будут благодарны за любую критику и конструктивные предложения по развитию математической модели русскоязычного текстового документа и готовы к сотрудничеству.

СПИСОК ЛИТЕРАТУРЫ

1. Сборник боевых документов Великой Отечественной Войны / ред. В. А. Небучинов, М. Н. Шарохин. – М. : Воениздат, 1947–1950. – Т. 1–12.

2. *Шелманов А. О.* Метод автоматического выделения многословных терминов из текстов научных публикаций // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16–20 октября 2012 г., г. Белгород, Россия): Труды конференции. Т. 1. – Белгород: Изд-во БГТУ, 2012. – С. 268–274.

3. *Шелманов А. О.* Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений / А. О. Шелманов, М. А. Каменская, М. И. Ананьева, И. В. Смирнов. // Искусственный интеллект и принятие решений. – 2016. – № 4. – С. 47–61.

4. *Шелманов А. О.* Извлечение информации из клинических текстов на русском языке / А. О. Шелманов, И. В. Смирнов, Е. А. Вишнёва // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). В 2 т. Т. 1. – М. : Изд-во РГГУ, 2015. – Вып. 14 (21). – С. 560–572.

5. *Суворов Р. Е.* Определение связанности научно-технических документов на основе характеристики тематической значимости / Р. Е. Суворов, И. В. Соченков // Искусственный интеллект и принятие решений. – 2013. – № 1. – С. 33–40.

6. *Клышинский Э. С.* Метод извлечения технических терминов с использованием меры странности / Э. С. Клышинский, Н. А. Кочеткова // Новые информационные технологии в автоматизированных системах. – 2014. – № 17. – С. 365–370.

7. *Браславский П. И.* Сравнение пяти методов извлечения терминов произвольной длины / П. И. Браславский, Е. А. Соколов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). – М. : Изд-во РГГУ, 2008. – Вып. 7 (14). – С. 67–74.

8. *Лукашевич Н. В.* Использование методов машинного обучения для извлечения слов-терминов / Н. В. Лукашевич, Ю. М. Логачев // XII национальная конференция по искусственному интеллекту с международным участием КИИ-2010 (20–24 сентября 2010 г., Тверь, Россия): Труды конференции. В 4-т. Т. 1. – М. : Физматлит, 2010. – С. 292–299.

9. *Шульга Т. Э.* О задаче автоматического извлечения терминов из текста / Т. Э. Шульга, А. С. Петров // Информационно-коммуника-

ционные технологии в науке, производстве и образовании ICIT-2016: материалы Международной научно-практической конференции, Саратов, 23–28 августа 2016 г. – Саратов : ООО Издательство «Научная книга», 2016. – С. 112–117.

10. *Жеребило Т. В.* Словарь лингвистических терминов / Т. В. Жеребило. – изд. 5-е, испр. и доп. – Назрань : Пилигрим, 2010. – 486 с.

11. *Захаров В. П.* Корпусная лингвистика: Учебник для студентов направления «Лингвистика» // В. П. Захаров, С. Ю. Богданова. – 2-е изд., перераб. и дополн. – СПб. : СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

12. *Frantzi K.* Automatic recognition of multi-word terms: The C-value/NC-value method / K. Frantzi, S. Ananiadou, H. Mima // International Journal on Digital Libraries. – 2000. – Т. 3. – № 2. – P. 115–130.

13. *Baroni M.* BootCaT: Bootstraping Corpora and Terms from the Web / M. Baroni, S. Bernardini // Proceedings of LREC 2004. – 2004. – Т. 4. – С. 1313–1316.

14. *Dagan I.* Termight: Identifying and translating technical terminology. / I. Dagan, K. Church // Proceedings of the fourth conference on Applied natural language processing. – 1994. – P. 34–40.

15. *Justeson J. S.* Technical terminology: some linguistic properties and an algorithm for identification in text / J. S. Justeson, S. M. Katz // Natural Language Engineering. – 1995. – № 1(1). – P. 9–27.

16. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages / M. Korobov // Analysis of Images, Social Networks and Texts. – 2015. – P. 320–332.

17. Словарь военных терминов / Сост. А. М. Плехов. – М. : Воениздат, 1988. – 335 с.

18. *Kozakov L.* Glossary extraction and utilization in the information search and delivery system for IBM Technical Support / L. Kozakov, Y. Park, T. Fin, Y. Drissi, Y. Doganata, T. Cofino // IBM Systems Journal. – 2004. – Т. 43 – № 3. – P. 546–563.

19. *Sclano F.* TermExtractor: a web application to learn the shared terminology of emergent web communities / F. Sclano, P. Velardi // Enterprise Interoperability II – New Challenges and Industrial Approaches, Proceedings of the 3th International Conference on Interoperability for Enterprise Software and Applications. – 2007. – P. 287–290.

Петров А. С. – аспирант кафедры Информационная безопасность автоматизированных систем, Институт прикладных информационных технологий и коммуникаций, Саратовский государственный технический университет им. Ю. А. Гагарина.
E-mail: p.a_saratov@mail.ru

Шульга Т. Э. – д-р физ.-мат. наук, профессор кафедры Информационно-коммуникационные системы и программная инженерия, Институт прикладных информационных технологий и коммуникаций, Саратовский государственный технический университет им. Ю. А. Гагарина.
E-mail: shulga@sstu.ru

Petrov A. S. – Aspirant, Department of Information Security of Automated Systems, Institute of Applied Information Technologies and Communications, Saratov State Technical University.
E-mail: p.a_saratov@mail.ru

Shulga T. E. – Doctor of Physico-Mathematical Sciencies, Professor, Department of Information and Communication Systems and Software Engineering, Institute of Applied Information Technologies and Communications, Saratov State Technical University.
E-mail: shulga@sstu.ru