УДК 004.912 DOI: 10.17587/it.26.30-38

Е. В. Полицына, канд. техн. наук, доц., e-mail: kathrin.beaver@mail.ru,
С. А. Полицын, канд. техн. наук, доц., e-mail: pul_forever@mail.ru,
А. О. Касаткина, студент, e-mail: alyona.kasatkina1997@yandex.ru,
Московский авиационный институт (национальный исследовательский университет)

Создание интегрального алгоритма и инструментов автоматического реферирования текстов на русском языке

В настоящее время количество информации, представленной в текстовом виде, с каждым годом увеличивается, тем самым задача автоматической обработки текстов, а особенно сокращения их объема, становится все более актуальной. Инструменты, предназначенные для получения реферата на русском языке, в основном используют только статистический метод. Существует необходимость исследования алгоритмов реферирования для создания новых алгоритмов и инструментов, которые будут предоставлять возможность использования нескольких методов автоматического реферирования для улучшения результатов. В статье представлены результаты исследования алгоритмов экстракции, на их основе предлагается интегральный алгоритм реферирования, библиотека и сервис автоматического реферирования текстов на русском языке для обеспечения возможности использования различных методов реферирования.

Ключевые слова: *экстракция, методы автоматического реферирования, интегральный метод реферирования, библиотека и сервис автоматического реферирования текста*

Введение

С развитием информационных технологий из года в год человечество накапливает все больше информации, в том числе представленной в текстовом виде. В связи с этим все более актуально создание автоматизированных средств ее анализа и обработки. Для удобства обработки больших массивов текстов часто необходимо получать из них наиболее важную информацию в краткой форме, обозримой для беглого просмотра или быстрого анализа.

С такой проблемой сталкиваются специалисты не только в области лингвистики, но и во многих других областях, например, редакторы и писатели при подготовке и анализе новостей, ученые при поиске информации по теме исследования, технические писатели и аналитики при подготовке документации и др. В повседневной жизни людей также постоянно возникает проблема избыточности получаемой информации, начиная от новостей из СМИ или от знакомых людей до результатов выдачи поисковых систем.

Средства автоматического построения краткого представления текста очень полезны в любых сферах, где людям необходимо справиться с огромными потоками информации и быстро принять решение о том, какая информация подходит для дальнейшей работы, а какую нужно отсеять на первом этапе. Для этого применяются методы реферирования — краткого изложения текста в письменном виде с раскрытием его основного содержания по всем затронутым вопросам [1].

О важности развития этого направления свидетельствуют и факты покупки в 2013 г. компаниями Google и Yahoo стартапов Wavii и Summly соответственно, которые занимались созданием и развитием средств реферирования. Компания "Яндекс" поддерживает работы в области автоматического реферирования веб-документов с учетом запроса [2]. Востребованность автоматических средств для сокращения объема текста подтверждают и все больше набирающие популярность мобильные приложения, которые обрабатывают новостные потоки и представляют пользователю короткие рефераты на выбранные

темы, а также создание плагинов для различных браузеров, которые позволяют получить краткое содержание веб-страниц.

Методы автоматического реферирования текстов и алгоритмы расчета весов предложений

Существует два подхода к автоматическому составлению реферата [3]: *экстракция* — извлечение из исходного текста наиболее важных информационных блоков (абзацев, предложений); *абстракция* — генерация реферата с порождением нового текста, содержательно обобщающего первичный документ. Второй подход опирается на построение модели понимания текста и его синтеза на естественном языке. Понимание смысла текста — это процесс, который происходит путем установления логических связей между предметами на основе имеющихся знаний [4], что является отдельной большой и нерешенной проблемой в компьютерной лингвистике. Таким образом, для применения подхода на основе абстракции необходимо построение семантической модели текста, наличие знаний о широком круге предметных областей в формализованном виде и разработка алгоритмов оперирования этими представлениями. Помимо сложности реализации и наличия ряда концептуальных проблем понимания текста человеком времени на получение реферата, таким образом, требуется существенно больше, чем для подхода на основе экстракции, а использование сложных структур данных и необходимость подготовки и хранения дополнительных исходных данных накладывает дополнительные требования к ресурсам компьютера.

Выделяются следующие методы экстракции [3]:

1) *статистический метод*, суть которого заключается в выделении в тексте частотных слов, вычислении весов предложений с помощью суммирования частот (весов) входящих в их состав слов и включении в реферат предложений с наибольшими весами;

2) *позиционный метод*, который опирается на предположение о том, что информативность текстового блока (предложения) находится в зависимости от его позиции (места) в тексте документа;

3) *индикаторный метод*, основанный на идентификации фраз первичного документа с помощью индексации их специальными словами — маркерами, индикаторами и коннекторами.

Помимо методов реферирования существуют также алгоритмы расчета веса предложения. Вес — количественная характеристика, отражающая значимость предложения в тексте. Для его определения выделяют следующие алгоритмы:

1. Статистический алгоритм, который основан на определении веса предложения в зависимости от весов входящих в него ключевых слов [5].

2. Алгоритм на основе симметричного реферирования, который заключается в вычислении связей данного предложения с другими [6].

Инструменты автоматического реферирования текстов

Помимо разработок компаний, занимающихся развитием поисковых систем или новостных агрегаторов, существует ряд инструментов автоматического реферирования текстов: систем с веб-интерфейсом, библиотек и т.д. Большинство из них поддерживают работу только с текстами на английском языке, наиболее интересными и работоспособными из мультиязычных систем являются SweSum, MEAD. Система MEAD представляет собой библиотеку на языке Python, графический интерфейс пользователя не предоставляется. Система SweSum предназначена для автоматического реферирования текстов на различных языках (в списке доступных языков нет русского языка). Система основана на трех методах реферирования: статистическом, позиционном и индикаторном. Даже несмотря на то, что в системе применяются три метода реферирования текста, в реферат не всегда включаются все предложения, отражающие смысл текста.

Построение рефератов по текстам на русском языке поддерживают следующие системы.

1. TextAnalyst

Система TextAnalyst используется для автоматического реферирования текстов на русском языке с помощью статистического метода, т. е. наиболее значимыми считаются те предложения, которые имеют наибольший вес. Главным недостатком программы является качество реферирования текста: выбранные предложения не полностью раскрывают смысл текста, а также не связаны друг с другом логически.

2. VisualWorld

В состав системы VisualWorld входит инструмент для автоматического реферирования текстов на русском языке "Рефератор", осно-

ванный на статистическом методе, т. е. в реферат включаются предложения с наибольшим весом. Главным недостатком системы является качество получаемого реферата: отобранные предложения не полностью раскрывают смысл текста и не всегда связаны друг с другом.

3. Text Summarization

Система Text Summarization предназначена для реферирования текстов на русском и английском языках с использованием статистического метода. Основным недостатком системы является качество получаемого реферата, так как отобранные предложения не связаны друг с другом логически.

4. TextCompactor

Система TextCompactor применяется для автоматического реферирования текстов на различных языках, в том числе и на русском языке. Система основана на статистическом и позиционном методах реферирования. Несмотря на то, что в системе применяются два подхода к реферированию текста, в реферат включаются не все предложения, отражающие основную мысль текста.

5. Tools4noobs

Система Tools4noobs применяется для автоматического реферирования текстов на различных языках, включая русский язык. Для построения реферата в системе используются статистический и позиционный методы. Главным недостатком системы является качество получаемого реферата: отобранные предложения не полностью передают смысл текста.

Таким образом, у инструментов, использующих более широкий спектр методов реферирования, нет поддержки русского языка, а инструменты, в которых имеется возможность реферирования текста на русском языке, используют только статистический или статистический и позиционный методы. Следовательно, существует необходимость в создании инструментов с возможностью реферирования текстов на русском языке, использующих более широкий набор методов реферирования.

Сравнительный анализ методов автоматического реферирования

Определение качества работы методов реферирования также является отдельным направлением исследования, так как отсутствуют какие-либо наборы эталонных рефератов. В ряде исследовательских работ [2] упоминаются подготовленные или полученные корпусы текстов новостей с их рефератами, но они не находятся в свободном доступе и содержат перефразированные предложения, передающие смысл исходного текста, что делает невозможным их применение для оценки качества реферирования методом экстракции.

Таким образом, для получения результатов работы описанных методов был подготовлен набор текстов на русском языке разных стилей — проведен опрос респондентов в целях получения эталонных рефератов. Эталонные рефераты — упорядоченные списки предложений по важности их в тексте на основании усредненной оценки группы респондентов разного возраста, профессии и т. д.

Были реализованы описанные методы автоматического реферирования: статистический, позиционный, индикаторный. Статистический метод реализован с применением статистического и симметричного алгоритмов расчета веса предложений. Индикаторный метод реализован с применением алгоритма на основе использования только индикаторов и алгоритма на основе использования маркеров, индикаторов и коннекторов.

В табл. 1 приведены результаты работы методов автоматического реферирования текстов в зависимости от стилей текста и объема реферата 30 %, 50 % и 70 %. В табл. 1 представлены результаты (в процентах — отношение предложений реферата полученного каким-либо методом к "эталонному") для двух разноплановых (хорошо структурированный и более описательный) текстов публицистического стиля, поскольку публицистическому стилю характерно применение разных способов изложения текста, развития мысли и т. д.

Анализ полученных результатов показал, что:

1) наиболее эффективный метод — статистический с применением симметричного алгоритма, в нем выбираются связанные друг с другом предложения, что позволяет достичь смысловой целостности реферата;

2) наименее эффективный метод — индикаторный (только индикаторы), так как для построения реферата недостаточно только находящихся в тексте индикаторов;

3) недостаток позиционного метода — отсутствие возможности получения реферата заданного объема;

4) недостаток индикаторного метода — большая зависимость от пользователя. Отсутствие маркеров и коннекторов уменьшает эффективность работы метода на 2,3—25,4 %;

Таблица 1

Результаты работы методов автоматического реферирования текстов

Стиль текста	Объем реферата, %	Статистический метод с применением статистического алгоритма	Статистический метод с применением симметричного алгоритма	Позиционный метод	Индикаторный метод	Индикаторный метод (только индикаторы)	Среднее значение
Публицистический	30	20	40		40	80	**46**
	50	45	56	50	78	67	**59,2**
	70	77	69		85	69	**70**
Среднее значение		**47,3**	**55**	**50**	**67,7**	**72**	
Публицистический	30	67	67		67	33	**59,6**
	50	64	64	64	64	55	**62,2**
	70	80	87		80	80	**78,2**
Среднее значение		**70,3**	**72,7**	**64**	**70,3**	**56**	
Научный	30	20	40		40	20	**43**
	50	63	63	75	63	25	**57,8**
	70	82	82		82	64	**77**
Среднее значение		**55**	**61,7**	**75**	**61,7**	**36,3**	
Художественный	30	42	42		42	33	**40,2**
	50	65	75	42	55	40	**55,4**
	70	71	82		71	64	**66**
Среднее значение		**59,3**	**66,3**	**42**	**56**	**45,7**	

5) с увеличением объема реферата увеличивается процент совпадений, что обусловлено увеличением числа ключевых слов и более полной передачей смысла текста;

6) наиболее эффективные результаты получаются при реферировании научных и публицистических текстов, а наименее — при реферировании художественных, так как для художественных текстов характерна неоднозначность их понимания, что отрицательно влияет на работу методов автоматического реферирования [7].

Алгоритмы на основе сочетания методов автоматического реферирования текстов

Проведенный сравнительный анализ работы методов автоматического реферирования текста показал, что каждый из них имеет определенные недостатки, многие из которых могут быть компенсированы дополнительным применением другого метода. Исходя из этого были разработаны алгоритмы на основе сочетания этих методов.

- *Алгоритм на основе сочетания позиционного метода со статистическим методом*

Недостатком позиционного метода является отсутствие возможности получения реферата заданного объема. При применении позиционного метода весь текст делится на ключевые предложения — те, которые находятся в начале и конце каждого абзаца, и не ключевые — остальные предложения текста. Все ключевые предложения имеют одинаковую значимость. Следовательно, становится невозможным увеличивать или уменьшать объем реферата.

Благодаря совместному использованию позиционного и статистического метода появляется критерий для оценки значимости предложения — его вес. Это значит, что объем реферата может быть изменен в зависимости от полученных весов предложений.

- *Алгоритм на основе сочетания индикаторного метода со статистическим методом*

Результаты работы индикаторного метода в большой степени зависят от человека, что является его основным недостатком этого метода. Реферат строится в зависимости от встречающихся в тексте маркеров, индикаторов и коннекторов, причем маркеры и коннекторы определяются человеком. Таким образом, если человек, применяющий этот метод, проигнорирует данную возможность или некорректно задаст маркеры и коннекторы, то вес предложений, который будет определяться только на основе находящихся в них индикаторов, будет сформирован неверно, что от-

рицательно скажется на результатах работы метода.

Применение сочетания индикаторного метода со статистическим методом позволяет использовать автоматически выделенные ключевые слова. Вес предложения в этом случае будет вычисляться на основе ключевых слов, выделяемых при использовании статистического метода, и индикаторов, используемых индикаторным методом.

Так как маркеры — ключевые слова и коннекторы — синонимы определяются человеком, то неправильное их выделение приведет к ухудшению работы методов. В результате опроса респондентов было выявлено, что задача выделения синонимов является непростой для большинства из них. Многие респонденты выделяют слова, основываясь на своих ассоциациях, которые при этом означают разные предметы, явления, действия, т. е. не являются синонимами, а находятся в отношении "род—вид" или связаны только ассоциативно. В результате реферат на основе индикаторного метода и алгоритма сочетания его со статистическим методом строится с использованием "ошибочных" синонимов, что приводит к ухудшению их работы. Таким образом, при реализации инструментов реферирования ввод синонимов человеком был исключен, впоследствии он может быть заменен использованием тематических словарей синонимов в сочетании с автоматическим определением предметной области текста [8, 9].

Интегральный метод автоматического реферирования текстов

На основе проведенного анализа полученных результатов применения алгоритмов реферирования текстов на основе сочетания различных методов был разработан интегральный метод автоматического реферирования текста.

Интегральный метод автоматического реферирования текста основан на комплексном использовании следующих методов:

- статистического с применением статистического алгоритма;
- статистического с применением симметричного алгоритма;
- позиционного;
- позиционного-статистического с применением статистического алгоритма;
- позиционного-статистического с применением симметричного алгоритма;
- индикаторного с использованием маркеров и индикаторов;
- индикаторного с использованием только индикаторов;
- индикаторного-статистического с применением статистического алгоритма;
- индикаторного-статистического с применением симметричного алгоритма.

Лежащий в основе интегрального метода алгоритм получения реферата состоит из следующих шагов:

1. Выделение ключевых предложений вышеописанными методами.

2. Расчет числа повторений каждого ключевого предложения:

$$k = \sum_{i=1}^{n} F_i(s), \qquad (1)$$

где s — ключевое предложение; $F_i(s)$ — функция, определяющая наличие ключевого предложения s во множестве выделенных i-м методом ключевых предложений; n — число методов;

$$F(s) = \begin{cases} 1, s \in M; \\ 0, s \notin M, \end{cases} \qquad (2)$$

где s — слово; M — множество методов выделений ключевых предложений.

3. Расчет порога пересечения:

- формирование числового ряда из полученных значений числа повторений всех ключевых предложений;
- определение моды полученного числового ряда p, p является порогом пересечения ключевых предложений.

4. Выделение предложений, число повторений которых больше или равно пороговому значению.

5. Сравнение числа полученных предложений с запрашиваемым объемом реферата:

- если число полученных предложений меньше, чем запрашиваемый объем реферата, то добавляются предложения, число повторений которых наиболее близко к порогу пересечения;
- если число полученных предложений больше, чем запрашиваемый объем реферата, то удаляются предложения с наименьшим числом повторений.

При получении ключевых предложений могут встречаться предложения с одинаковым числом повторений. В этом случае невозможно определить, какое из них следует удалить или добавить, поэтому используется одно, выбранное случайным образом предложение.

6. Сортировка полученных предложений по их номерам и получение итогового списка ключевых предложений и реферата.

В табл. 2 приведены результаты сравнения работы алгоритмов интегрального метода.

Анализ полученных результатов показал, что:

1) применение интегрального метода с использованием маркеров позволяет получить более эффективные результаты, чем без их использования, так как реферат строится на основе автоматически выделенных ключевых слов и введенных пользователем маркеров;

2) использование интегрального метода с применением алгоритма случайного выбора предложений с использованием маркеров позволяет получить более эффективные результаты, чем без их использования, так как реферат строится на основе автоматически выделенных ключевых слов и введенных пользователем маркеров;

3) использование интегрального метода с применением алгоритма случайного выбора предложений позволяет получать каждый раз разный реферат, что предоставляет пользователю возможность выбора наиболее подходящего из них;

4) в некоторых случаях использование интегрального метода с применением алгоритма случайного выбора предложений позволяет получить более эффективные результаты, чем использование "чистого" интегрального метода, так как выбираются предложения, число повторений которых меньше порогового значения, но которые являются ключевыми предложениями по мнению респондентов.

В табл. 3 приведены результаты сравнения работы интегрального метода и других существующих инструментов автоматического реферирования.

По данным табл. 3 можно сделать вывод, что использование интегрального метода в большинстве случаев позволяет получить более эффективные результаты (в среднем на 15,7 %), чем использование других инструментов реферирования.

На рис. 1 представлен график зависимости времени построения реферата от объема исходного текста. Предложенный интегральный алгоритм имеет линейную сложность, что позволяет использовать его в качестве одного из

Таблица 2

Результаты сравнения работы алгоритмов интегрального метода

Стиль текста	Объем реферата, %	Интегральный метод	Интегральный метод (случайный выбор)	Интегральный метод (без маркеров)	Интегральный метод (случайный выбор, без маркеров)
Публицистический	30	40	38,4	20	33,7
	50	67	56,4	56	50
	70	77	68,3	69	64,7
Среднее значение		61,3	54,4	48,3	51,5
Публицистический	30	67	58,8	67	47,8
	50	73	66,8	73	62,5
	70	87	74,4	87	69,8
Среднее значение		75,7	66,7	75,7	60,1
Научный	30	40	43,2	40	36,3
	50	75	61,7	63	53,7
	70	82	76,9	82	79,4
Среднее значение		65,7	60,6	61,7	56,5
Художественный	30	33	31,5	33	30,3
	50	60	63,3	55	58,7
	70	75	74,2	75	76,2
Среднее значение		56	56,3	54,3	**55,1**

Таблица 3

Результаты сравнения работы интегрального метода и других существующих инструментов автоматического реферирования

Стиль текста	Объем реферата, %	Интегральный метод	Text Summarization	Tools4noobs
Публицистический	30	40	40	60
	50	67	67	67
	70	77	77	69
Среднее значение		**61,3**	**61,3**	**65,3**
Публицистический	30	67	17	50
	50	73	45	55
	70	87	60	73
Среднее значение		**75,7**	**40,7**	**59,3**
Научный	30	40	20	20
	50	75	50	63
	70	82	64	82
Среднее значение		**65,7**	**44,7**	**55**
Художественный	30	33	25	17
	50	60	45	40
	70	75	71	61
Среднее значение		**56**	**47**	**39,3**

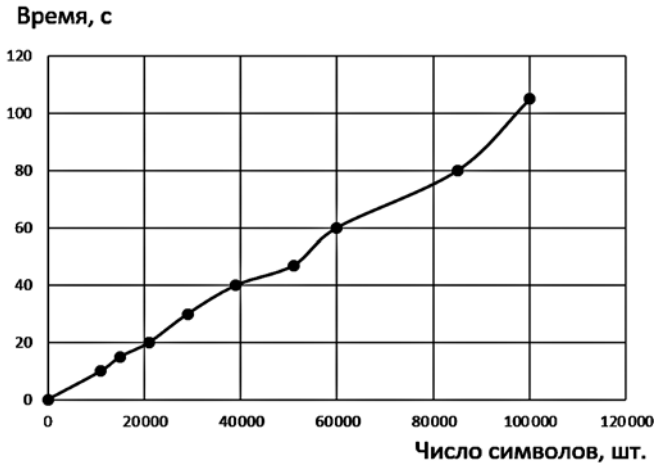

Рис. 1. График зависимости времени построения реферата от объема исходного текста

этапов при обработке больших наборов текстов, в том числе в прикладных системах.

Библиотека автоматического реферирования текста

С использованием всех реализованных методов была разработана Java-библиотека Summarization [10] для обеспечения возможности использования этих методов в других системах и программных комплексах.

Входящие в разработанную библиотеку методы делятся на две группы [10].

В первую группу входят методы, предназначенные для работы с текстом:

1. Метод для получения всех слов текста.
2. Метод для получения всех ключевых слов текста интегральным методом [11].
3. Метод для разделения текста на абзацы.
4. Метод для разделения абзацев на предложения.
5. Метод для получения слов каждого предложения в форме, в которой они употребляются в тексте.
6. Метод для получения слов каждого предложения в начальной форме.
7. Метод для получения ключевых слов каждого предложения.

Во вторую группу входят методы, предназначенные для получения реферата вышеописанными методами, и метод для получения списка предложений без их номеров в исходном тексте.

Сервис автоматического реферирования текста

С использованием разработанной библиотеки был создан веб-сервис с графическим интерфейсом пользователя и программным REST-интерфейсом, что позволяет использовать реализованные методы реферирования как пользователями, так и сторонними системами [10].

Сервис автоматического реферирования текста доступен на портале "Автоматизированный анализ текста" по адресу: http://abstracts.textanalysis.ru/, он предоставляет возможность получения реферата и ключевых слов. Для этого необходимо выполнить следующие шаги:

1. Вставить исходный текст в текстовое поле.
2. Выбрать метод реферирования. Все методы разделены на две группы в зависимости от того, необходимо ли наличие маркеров для применения метода (рис. 2).
3. Указать объем реферата в процентах.
4. Ввести маркеры, если выбран соответствующий метод.
5. Нажать кнопку "Получить реферат!".

В результате обработки ответа полученный реферат и ключевые слова отобразятся в соответствующих текстовых полях (рис. 3).

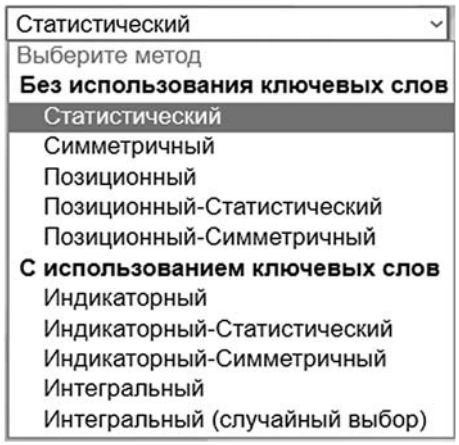

Рис. 2. Выпадающий список для выбора метода реферирования

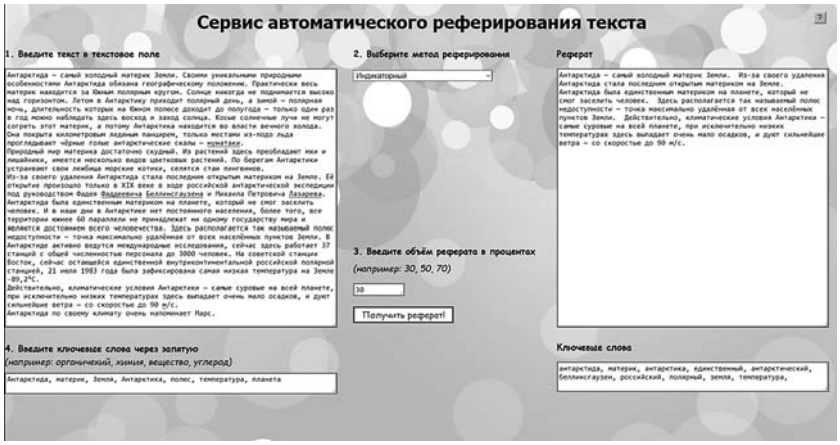

Рис. 3. Пример полученного реферата

Если пользователь не выбрал метод реферирования и дополнительные настройки, то будет автоматически применен интегральный метод реферирования текста.

Также пользователь может получить справочную информацию по использованию сервиса.

На основе полученных отзывов пользователей сформированы направления развития сервиса реферирования:

- определение числа символов, слов, предложений, абзацев по вводимым текстам и получаемым рефератам;
- создание возможности задания объема реферата не только в процентах, но в числе предложений и символов;
- создание возможности выбора предложений, которые должны быть включены в реферат;
- разбиение получаемого реферата на абзацы в соответствии с абзацами исходного текста;
- выделение ключевых слов в получаемом реферате.

Заключение

Разработанный интегральный алгоритм реферирования позволяет получать более точные по сравнению с эталонным рефератом результаты, чем другие существующие инструменты. Помимо развития сервиса реферирования дальнейшими направлениями исследования для совершенствования методов реферирования являются:

1. Использование ключевых словосочетаний и именованных сущностей в дополнение к ключевым словам при применении методов автоматического реферирования.

2. Дальнейшее исследование методов реферирования, в том числе дополнение интегрального метода распространенными алгоритмами для английского языка TextRank, LexRank, LSA.

3. Проведение дополнительного анализа характеристик текста (стиль, объем) с последующим автоматическим выбором наиболее подходящего метода.

4. Использование инструментов для устранения орфографических и пунктуационных ошибок, расшифровки сокращений в исходном тексте.

5. Использование подготовленных по разным тематикам и стилям текстов словарей синонимов [8] при применении индикаторного метода и алгоритмов его сочетания со статистическим методом.

6. Выделение определенного числа ключевых слов в зависимости от объема текста.

Библиотека и сервис автоматического реферирования текста предоставляют возможность использования не только существующих методов автоматического реферирования, а также алгоритмов на основе их сочетания и интегральный метод. С помощью разработанного сервиса были получены рефераты по текстам различных стилей и объемов и разным видам технической документации, программный интерфейс сервиса использован в мобильном приложении Tourist Helper 2.0 [12].

Сервис полезен для специалистов, работающих с большим объемом текстовых данных, и позволяет быстро получить краткий вариант текста нужного объема. Наличие программного интерфейса сервиса и библиотеки в свободном доступе дает возможность автоматизации сокращения текстовых данных в рамках любого процесса.

Список литературы

1. **Маркушевская Л. П., Цапаева Ю. А.** Аннотирование и реферирование: методические рекомендации для самостоятельной работы студентов. СПб.: ИТМО, 2008. 8 с.

2. **Браславский П. И., Колычев И. С.** Автоматическое реферирование веб-документов с учетом запроса // Интернет-Математика 2005. Автоматическая обработка веб-данных. М.: Яндекс, 2005. С. 485—501.

3. **Тарасов С. Д.** Современные методы автоматического реферирования // Научно-технические ведомости СПбГПУ. 2010. № 6. С. 59—74.

4. **Букатникова С. Д.** Понимание текста как проблема современной лингвистики и гуманитарного познания // Молодой ученый. 2015. № 6. С. 799—803.

5. **Лазарева О. Ю., Боломутова М. С.** Методы выделения ключевых слов в контексте электронных обучающих систем // Молодой ученый. 2016. № 22. С. 143—146.

6. **Яцко В. А.** Симметричное взвешивание терминов // Символ науки. 2015. № 12. С. 87—88.

7. **Позняк Г. В.** Психолингвистические основы реферирования как учебной деятельности // Факультет международных отношений БГУ: электронный сборник. Вып. II. Минск: БГУ, 2012. С. 33—38.

8. **Милованова Е. Е.** Определение семантической схожести текстов информационными системами // Гагаринские чтения. 2019: XLV Международная молодежная научная конференция: Сборник тезисов докладов. М.: Московский авиационный институт (национальный исследовательский университет), 2019. С. 381.

9. **Белов С. М.** Создание программной системы классификации текстов // Материалы XVIII Международной конференции "Информатика: проблемы, методология, технологии". Секция 10: компьютерная лингвистика. 2018. Т. 6. С. 8—12.

10. **Документация** Java-библиотеки Summarization и сервиса автоматического реферирования текста. URL: http://textanalysis.ru/jce/details/our-instr, свободный. (Дата обращения: 14.06.19).

11. **Иващенко М. В.** Анализ методов автоматизированного выделения ключевых слов из текстов на естественном языке // Материалы XVIII Международной конференции "Информатика: проблемы, методология, технологии". Секция 10: компьютерная лингвистика. 2018. Т.6. С. 19—24.

12. **Приложение** Tourist Helper 2.0 / Google Play. URL: https://play.google.com/store/apps/details?id=ru.textanalysis.touristhelper, свободный (дата обращения: 20.06.19).

E. V. Politsyna, Cand. of Tech. Sc., Associate Professor, e-mail: kathrin.beaver@mail.ru,
S. A. Politsyn, Cand. of Tech. Sc., Associate Professor, e-mail: pul_forever@mail.ru,
A. O. Kasatkina, Student, e-mail: alyona.kasatkina1997@yandex.ru,
Moscow Aviation Institute (National Research University), Moscow, Russian Federation

Development of Integrated Algorithm and Tools of Automatic Summarization for Texts in the Russian Language

The amount of information provided in text form is increasing every year. Thus, the task of automatic natural language text processing (NLP), and especially the reduction of its volume, is becoming increasingly important. But tools which produce abstracts in the Russian language mainly use the statistical method only. So, there is a need to research the summarization algorithms and create new algorithms and tools that will use several automatic summarization methods simultaneously to improve the results. The article shows the results of extraction algorithms, basing of which an integrated summarization algorithm, a library and service of automatic text summarization in the Russian languages are proposed to enable the use of various reference methods.

Keywords: *data extraction, methods of summarization, integrated summarization method, automatic summarizations service*

DOI: 10.17587/it.26.30-38

References

1. **Markushevskaya L. P., Tsapaeva Y. A.** Annotation and summarization: guidelines for students' self-studies, Spb., Publishing house of ITMO, 2008, 8 p. (in Russian).

2. **Braslavsky P. I., Kolychev I. S.** Automated summatization of web-ducuments basing on search requests, In proceedings: Internet-matematika 2005. Automated web-data processing. Moscow, Yandex, pp. 485—501 (in Russian).

3. **Tarasov S. D.** Modern methods of automated summarization, *Nauchno-technicheskie vedomosti SpbGpu*, 2010, no. 6, pp. 59—74 (in Russian).

4. **Bukatnikova S. D.** Text understanding as a modern problem of linguistics and humanitarian knowledge, *Molodoy Ucheniy*, 2015, no. 6, pp. 799—803 (in Russian).

5. **Lazareva O. Y., Bolotoumova M. S.** Methods of keyword extraction in computer educational systems, *Molodoy Ucheniy*, 2016, no. 22, pp. 143—146 (in Russian).

6. **Yatsko V. A.** Symmetric term weighting, Simvol nauki, 2015, no. 12, pp. 87—88 (in Russian).

7. **Posnyak G. V.** Psycholinguistic basics of summarizing as a learning, *Faculty of International Affairs BGU:* vol. II, Minsk, BGU, 2012, pp. 33—38 (in Russian).

8. **Milovanova E. E.** Determination of semantic similarity of texts, *XLV Mezhdunarodnaya molodezhnaya nauchnaya konferenciya: sbornik dokladov*, Moscow, MAI, 2019, p. 381 (in Russian).

9. **Belov S. M.** Development of the text classification system, *Materialy XVIII Materials of XVIII International Conference "Informatika: problemy, metodologiya, tekhnologii",* vol. 6, pp. 8—12 (in Russian).

10. **Summarization** library Java-docs and automated summarization service documentation, available at: http://textanalysis.ru/jce/details/our-instr (access date: 14.06.19).

11. **Ivashchenko M. V.** Analysis of automated keyword extraction methods in natural language texts, *Materials of XVIII International Conference "Informatika: problemy, metodologiya, tekhnologii". Sec. 10: computer linguistics*, 2018, vol. 6, pp. 19—24 (in Russian).

12. **Application** Tourist Helper 2.0, Google Play, available at: https://play.google.com/store/apps/details?id=ru.textanalysis.touristhelper. — (access date: 20.06.19).