УДК 004.048:519.765
DOI 10.25205/1818-7900-2018-16-3-74-86

Т. В. Батура [1, 2], **А. М. Бакиева** [1]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Институт систем информатики им. А. П. Ершова СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

tatiana.v.batura@gmail.com, m_aigerim0707@mail.ru

СОЗДАНИЕ СИСТЕМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ

Описан новый метод автоматического реферирования текстов. На основе предложенного метода создана система, позволяющая получать краткие аннотации научно-технических текстов и определять их темы. Процесс реферирования состоит из пяти основных шагов: предобработка, риторический анализ и преобразование текста, оценка весов, выбор предложений и сглаживание. Предлагаемый метод формирует аннотацию на основе наиболее значимых предложений исходного документа. Значимость предложений частично определяется в процессе риторического анализа, который выполняется с помощью дискурсивных маркеров и коннекторов. Также учитываются ключевые слова, многословные термины и некоторые специальные слова, которые часто встречаются в научно-технических текстах. Для извлечения ключевых слов и определения тем текста применялась аддитивная регуляризация тематических моделей.

Ключевые слова: автоматическое реферирование, теория риторических структур, дискурсивные маркеры, аддитивная регуляризация, тематические модели.

Введение

Ввиду стремительного увеличения объемов текстовой информации в Интернете активные исследования в области компьютерной лингвистики сохраняют свою актуальность. Разработка алгоритмов и создание систем автоматического реферирования, поиска и извлечения информации, классификации и кластеризации текстовых документов по-прежнему являются сложными задачами.

Подход, основанный на применении дискурсивного анализа, используется довольно широко для решения различных задач компьютерной лингвистики. Подробный обзор литературы, представленный в работе [1], показывает, что в большинстве случаев дискурсивный анализ способен улучшить качество автоматических систем на 4–44 % в зависимости от конкретной задачи.

В работе [2] теория риторических структур применяется для определения важных предложений в документе. Автор представляет входной текст в виде набора деревьев и предлагает использовать алгоритм ограничений для объединения этих деревьев. Далее применяется несколько эвристик для выбора более подходящих деревьев при формировании реферата. Автоматизированная многоязычная система реферирования текста SUMMARIST описана в [3]. Эта система сочетает в себе методы понятийного уровня знаний о мире, методы информационного поиска и статистические методы. Алгоритм состоит из трех этапов: иденти-

Батура Т. В., *Бакиева А. М.* Создание системы автоматического реферирования научных текстов // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 74–86.

фикация темы, интерпретация и генерация. SUMMARIST формирует аннотации на пяти языках: английском, японском, испанском, индонезийском и арабском.

Система автореферирования научных статей, основанная на дискурсивном анализе, описана в [4]. В ней определены семь риторических категорий. Автор работы [5] применил теорию риторических структур для создания графического представления документа. На основе структурного анализа текста вычисляются веса предложений, из которых в итоге получается краткая аннотация. В работе [6] обсуждается создание реферата, содержащего не только информацию из одного конкретного документа, но и дополнительные знания из других, похожих на него по тематике документов.

С. Митхун описывает подход, базирующийся на схемах, для формирования аннотаций на основе запросов, в которых используются структуры дискурса [7]. Этот подход выполняет четыре основные задачи, а именно: категоризация вопроса, идентификация риторических предикатов, выбор схемы и обобщение. Автор создал систему BlogSum и оценил ее производительность относительно релевантности и согласованности вопросов. Полученные результаты показывают, что предлагаемый подход решает проблему несоответствия и дискурсивной несогласованности автоматически созданных рефератов.

Исследования в этой области для английского языка достигли достаточно высокого уровня, но для текстов на русском языке данная область изучена сравнительно мало. Анализ подходов для решения проблемы автоматического формирования рефератов научно-технических текстов на русском языке проводился российскими учеными в работах [8; 9]. В исследовании [8] описаны методы и алгоритмы, учитывающие нелинейный и иерархический характер текста. С помощью риторических отношений решается проблема экстракции (извлечения фрагментов текста). С. А. Тревгода разработал систему, основанную на правилах вывода и узкоспециализированном словаре, ключевых фраз. Гибридный подход, предложенный П. Г. Осмининым [9], сочетает методы экстракции и абстракции. Этот подход был реализован автором в системе реферирования, ориентированной на автоматический перевод. Описанная система построена для текстов по теме «математическое моделирование». Были использованы не только риторические структуры, но и глаголы из предметной области «математической логики». С помощью найденных ключевых слов определяется вес предложения, затем полученная аннотация формируется в соответствии с шаблонами.

Некоторые особенности риторических отношений описаны в работах [10; 11]. Там также формулируются утверждения о свойствах этих признаков. Работа [12] описывает опыт построения корпуса на русском языке, содержащего дискурсивные маркеры. Корпус общедоступен, включает в себя тексты разных жанров, таких как научный, научно-популярный, новостной. Прежде чем использовать теорию риторических структур, ее приходится адаптировать для конкретного языка. Это связано с грамматическими особенностями. В своей статье авторы предлагают иерархию риторических отношений, которая, согласно их исследованиям, является наиболее удобной и корректной для работы с текстами на русском языке.

В нашей работе описывается подход, позволяющий получать краткие аннотации научно-технических текстов и определять их темы. Предлагаемый метод формирует аннотацию на основе наиболее значимых предложений исходного документа. Значимость предложений частично определяется в процессе риторического анализа. Для определения тем текстов применяется метод аддитивной регуляризации тематических моделей (АРТМ) [13]. Этот метод позволяет решить проблему неединственности и неустойчивости при помощи введения дополнительных ограничений на требуемое решение. В качестве регуляризаторов могут использоваться: сглаживание и разреживание распределений терминов в темах, сглаживание и разреживание распределений тем в документах и др.

Семантический анализ и особенности риторических отношений

Теория риторических структур – одна из наиболее широко используемых теорий организации текстов [14]. Согласно ей, изначально текст делится на неперекрывающиеся фрагменты, а именно на элементарные дискурсивные единицы (ЭДЕ). Кроме того, последовательные ЭДЕ связаны риторическими отношениями. Эти части известны как элементы, из которых

строятся более крупные фрагменты текстов и целые тексты. Каждый фрагмент по отношению к другим фрагментам выполняет определенную роль. Текстовая связь формируется с помощью тех отношений, которые моделируются между фрагментами в тексте.

В теории риторических структур можно определить два типа ЭДЕ. Один из них, называемый ядром, считается наиболее важной частью высказывания, другой, называемый сателлитом, поясняет ядро и считается вторичным. Ядро содержит основную информацию, тогда как сателлит содержит дополнительную информацию о ядре. Сателлит часто непонятен без ядра. Между тем выражения, в которых сателлит удален, могут быть поняты лишь в некоторой степени. Рассмотрим следующий пример:

Текст: *Дом выглядел неплохо.* ***Кроме того***, *цена была подходящая.*
Маркер: *Кроме того*
Название отношения: Elaboration

Для удобства введем следующие обозначения:

x – ядро; y – маркер; z – сателлит;
$S(x)$ – предикат для ЭДЕ, которая является ядром;
$S'(x)$ – предикат для ядра, которое начинается с прописной буквы, т. е. находится в начале предложения;
$S(z)$ – предикат для ЭДЕ, которая является сателлитом;
$S'(z)$ – предикат для сателлита, который начинается с прописной буквы;
y' – маркер с прописной буквы;
$p(\ )$ – символ пунктуации, аргументом может быть ".", ",", ":", ";".

Теперь приведенный пример может быть представлен в виде формулы исчисления предикатов: $S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.)$.

Формальное описание преобразования текста

В предлагаемом подходе риторический анализ используется на этапе построения квазиреферата. Под квазирефератом понимается перечень наиболее значимых предложений текста. Упрощенно этот этап можно описать следующим образом. Сначала необходимо найти в тексте ядерные ЭДЕ. Далее следует преобразовать высказывания, содержащие эти ЭДЕ, так, чтобы получился сокращенный текст, являющийся промежуточным между исходным текстом и готовой аннотацией. В зависимости от разных маркеров и дискурсивных отношений эти преобразования будут разными. Для формального описания действий, выполняемых системой, было принято решение использовать логику предикатов первого и второго порядков.

Предикаты первого порядка

Согласно обозначениям, введенным в предыдущем разделе, для рассмотренного примера действия, выполняемые системой на этом этапе, могут быть записаны в таком виде:

$$S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.) \rightarrow S'(x) \wedge p(.) \wedge \neg(y' \wedge S(z) \wedge p(.)).$$

А именно, вначале надо найти маркер y = «*кроме того*», потом необходимо удалить его вместе с сателлитом, оставив предыдущее предложение, которое является ядерным ЭДЕ.

Для предикатов, представленных выше, мы ввели специальные *действия*, которые выполняются для создания квазиреферата. Они зависят от некоторых глаголов, существительных, маркеров и коннекторов.

Маркеры (*дискурсивные маркеры*) – это слова или фразы, которые не имеют реального лексического значения, но вместо этого обладают важной функцией формировать разговор-

ную структуру, передавая намерения говорящих. Примеры маркеров и действий, связанных с ними, приведены в табл. 1.

Коннекторы – группы слов, заменяющие маркеры и характеризующие определенные риторические отношения. Коннекторы обеспечивают связь между фразами, они показывают семантическую неполноту предложения. Например, «*в связи с этим*», «*вместе с тем*», «*тем самым*» и т. д. (табл. 2).

Таблица 1

Действия для маркеров

№	Риторические отношения	Маркеры	Действия
1	Elaboration	Кроме того	SAVE_DELETE
2	Cause-Effect	Поэтому	DELETE_SAVE
3	Contrast	Однако	SAVE_DELETE
4	Restatement	Другими словами	SAVE_DELETE
5	Elaboration	Например	SAVE_DELETE
6	Evidence	Таким образом	DELETE_SAVE

Таблица 2

Действия для коннекторов

№	Риторические отношения	Коннекторы	Действия
1	Elaboration	В связи с этим	SAVE_SAVE
2	Elaboration	Вместе с тем	DELETE_SAVE
3	Elaboration	Тем самым	SAVE_SAVE

Во время исследования мы создали словарь, состоящий из 121 маркеров и коннекторов, 120 существительных и 108 глаголов с весами, которые часто встречаются в научных и технических текстах. Всего было рассмотрено восемь действий. Ниже описаны некоторые действия.

DELETE_SAVE – это действие удаляет предыдущее предложение и сохраняет предложение с заданным маркером.

SAVE_DELETE – это действие сохраняет предыдущее предложение и удаляет предложение с заданным маркером.

SAVE_SAVE – это действие полностью сохраняет предложение с заданным маркером и предыдущим предложением.

Предикаты второго порядка

Как известно, в сложноподчиненном предложении выделяются главное и придаточное предложение. В этом случае ЭДЕ более низкого уровня вложены в ЭДЕ более высокого уровня. Для описания действий с вложенными ЭДЕ удобнее использовать предикаты второго порядка. Чтобы проиллюстрировать, как текст преобразуется в случае вложенных ЭДЕ, приведем следующий пример:

Кроме того, воздух, который поступает в морозильную камеру, уже охлаждают до 1.5 °C с помощью холодильной установки док-станции, которая составляет около 50 % тепловой нагрузки поступающего воздуха. **Таким образом**, чистым эффектом охлаждаемой док доставки является уменьшение инфильтрации нагрузки. Чистая прибыль равна разнице между уменьшением инфильтрации нагрузки морозильной камеры и холодильной нагрузки в доке судоходства. **Обратите внимание**, что док-холодильники работают при значительно более высоких температурах (1.5 вместо –23 °C) и потребляют значительно меньше энергии на ту же сумму охлаждения.

Для того чтобы записать преобразования в формальном виде, добавим следующие обозначения:

m – ядро в придаточном предложении;
n – сателлит в придаточном предложении;
$S(m)$ – предикат для ядра m;
$S(m)$ – предикат для ядра m, начинающегося с прописной буквы;
$S(n)$ – предикат для сателлита n;
$S'(n)$ – предикат для сателлита n, начинающегося с прописной буквы;
y – маркер.

$$S'(z) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(z) \wedge p(.) \rightarrow$$
$$\neg S'(El \wedge p(,) \wedge S(m) \wedge p(.)) \wedge \neg S'(Ev \wedge p(,)) \wedge S'(S(m) \wedge p(.)) \wedge S'(x) \wedge p(.) \wedge$$
$$\neg S'(Cont \wedge S(n) \wedge p(.)),$$

где
$y = El =$ «*кроме того*»;
$y = Ev =$ «*таким образом*»;
$y = Cont =$ «*обратите внимание*».

Следует отметить, что использование формализмов логики первого и второго порядка с данной целью пока недостаточно исследовано. В будущем, возможно, придется дополнить этот формализм, чтобы учитывался порядок следования элементов в тексте.

Общее описание системы

Пусть T – текст статьи, очищенный после предварительной обработки и состоящий из предложений

$$T = \bigcup_{k=1}^{P} S_k.$$

$D = \{d_1, d_2, ..., d_N\}$ представляет собой набор дискурсивных маркеров и коннекторов, которые содержатся в этом тексте.

$V = \{v_1, v_2, ..., v_M\}$ представляет собой набор глаголов и существительных, которые часто встречаются в научных и технических текстах.

В нашем понимании задача реферирования состоит в том, чтобы найти преобразование текста T в реферат $\tilde{T}$ такое, что $\Phi: T \rightarrow \tilde{T}$, $|\tilde{T}| < |T|$. Тогда алгоритм построения реферата можно записать в виде последовательных этапов.

1. Предобработка текста. На этапе предварительной обработки из исходного текста удаляются все изображения, таблицы, предложения с формулами, информация об авторах и библиографические ссылки. Авторские аннотации были убраны и отдельно сохранены, чтобы потом можно было оценить систему, путем сравнения результата с исходной аннотацией.

2. Риторический анализ и преобразование текста. На этом шаге обнаруживаются предложения, содержащие дискурсивные маркеры и коннекторы. К этим предложениям применяются определенные действия (см. выше). В результате получается квазиреферат: $\Phi(T, D, V) = T'$.

3. Оценка весов предложений. Для формирования аннотации подсчитываются веса предложений. Опишем эту процедуру подробнее.

Пусть S'_k – произвольное предложение квазиреферата

$$T' = \bigcup_{k=1}^{P_1} S'_k.$$

При вычислении веса каждого предложения квазиреферата учитывается наличие в этом предложении ключевых слов (или многословных терминов), дискурсивных маркеров и коннекторов, а также некоторых слов, которые характерны для научных текстов. Для извлечения из текстов многословных терминов используется алгоритм Turbotopics, разработанный для определения значимых *n*-грамм в английских текстах [15]. В ходе создания системы мы адаптировали алгоритм Turbotopics для работы с текстами на русском языке.

В итоге вес каждого предложения вычисляется по следующей формуле:

$$SW = \frac{1}{L} \cdot \sum_{i=1}^{L} w_i + \frac{1}{M} \cdot \sum_{j=1}^{M} v_j + \frac{1}{N} \cdot \sum_{k=1}^{N} d_k,$$

где

$W = \{w_1, ..., w_l\}$ – множество ключевых слов и многословных выражений $(|W| = L)$;

$V = \{v_1, ..., v_M\}$ – множество значимых глаголов и существительных, которые часто встречаются в научных текстах $(|V| = M)$;

$D = (d_1, ..., d_N)$ – множество дискурсивных маркеров и коннекторов $(|D| = N)$.

4. Выбор предложений. Из полученного набора предложений (см. п. 2) для аннотации отбираются только те, вес которых (см. п. 3) превышает заданную пороговую величину β:

$$\tilde{T} = \bigcup_{k=1}^{N_1} \{S'_k : SW > \beta\},$$

где β = 0,15 является константой, которая определяется эмпирически.

5. Операция сглаживания – процедура преобразования текста, позволяющая получить связный текст из разрозненных фрагментов и при необходимости дополнительно сократить его. Например, в процессе сглаживания заменяются или удаляются некоторые слова или словосочетания и т. д. В табл. 3 приведены примеры предложений до сглаживания и после него.

Таблица 3

Примеры сглаживания

До сглаживания	После сглаживания
Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ, **так как** в некоторых ситуациях эпизоды могут быть настолько продолжительными, что задержки процесса обучения, связанные с необходимостью завершения эпизодов, будут слишком велики.	Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ.
Поскольку, **как уже отмечалось**, использование БСД позволяет анализировать лишь один из возможных диагнозов.	**Выявлено что**, использование БСД позволяет анализировать лишь один из возможных диагнозов.

Для сглаживания предложений используются шаблоны и индикаторы. Мы рассмотрели два типа шаблонов: для удаления фрагментов предложений (в случае когда аннотация получилась длиннее 250 слов) и для дополнения (в случаях, когда в аннотацию попал фрагмент незаконченного предложения).

Индикаторы – комбинации слов, влияющие на вес предложения. В состав индикаторов входят определенные значимые слова, которые мы включили в созданную лингвистическую базу знаний. Примеры индикаторов и действий, связанных с ними, представлены в табл. 4. В данный момент рассмотрено 95 индикаторов.

Таблица 4

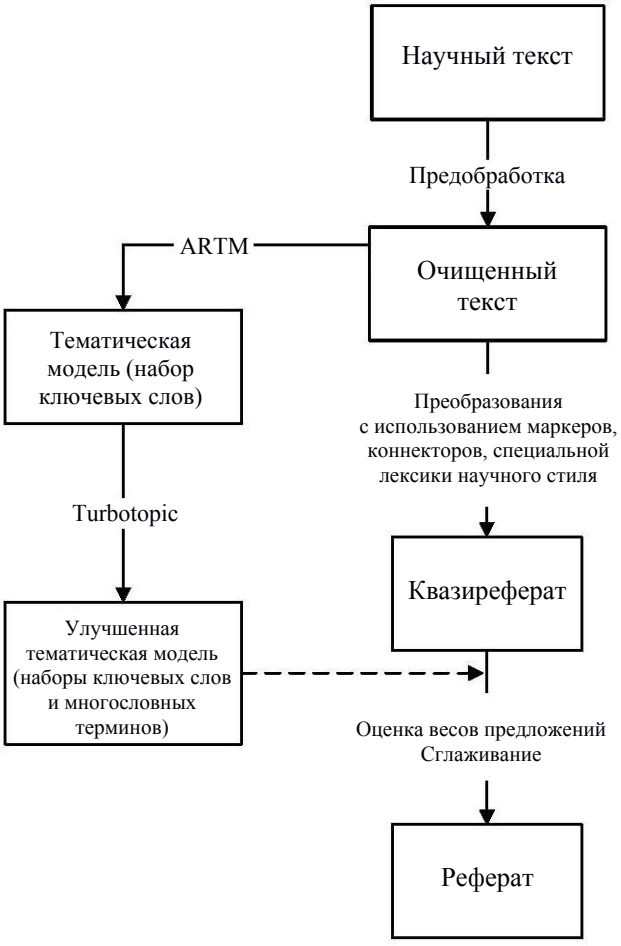
Действия для индикаторов

№	Индикатор	Действие	Результат
1	В нашей статье	REPLACE	В статье
2	Существенным является	DELETE_UNTIL_DOT	–
3	Следует подчеркнуть	DELETE_COMMA	–
4	Важным представляется	REPLACE	–

Действия, используемые при сглаживании:

REPLACE – замена индикатора на другое слово или словосочетание, или удаление данного индикатора;
DELETE_UNTIL_DOT – удаление до конца предложения, начиная с данного индикатора;
DELETE_COMMA – удаление фрагмента предложения до следующей запятой;
DELETE – удаление всего предложения.

В ходе данной работы была разработана система, блок-схема которой представлена далее:



Тематическое моделирование заключается в построении модели некоторой коллекции текстовых документов. В такой модели каждая тема представляется дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем.

В настоящее время существуют разные методы тематического моделирования, такие как PLSA, LDA, ARTM. Главное преимущество тематических моделей в сравнении с нейронными сетями заключается в том, что они хорошо поддаются интерпретации, пользователю понятны причины обнаружения определенных тем в тексте и структура самих тем. Кроме того, часто требуется, чтобы тематические модели учитывали разнородные данные, выявляли динамику тем во времени, автоматически разделяли темы на подтемы, использовали не только отдельные ключевые слова, но и многословные термины и т. д.

Чтобы выбрать алгоритм тематического моделирования, мы провели ряд экспериментов, результаты которых представлены в работе [16]. Было принято решение использовать алгоритм ARTM в реализации библиотеки BigARTM [17]. Благодаря своей универсальности и гибкости настройки параметров моделей ARTM позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. Этот метод гарантирует единственность и устойчивость решения. У ARTM не наблюдается увеличение количества параметров модели с ростом числа документов, поэтому он может применяться к большим наборам данных. Кроме того, предложенная нами модификация позволяет использовать не только однословные, но и многословные выражения, что, на наш взгляд, повышает интерпретируемость модели.

Результаты

Наша система была протестирована на коллекции из 261 научной статьи на русском языке. Эта коллекция собрана на основе выложенных в открытом доступе архивов журналов «Программные продукты и системы» за 2013–2017 гг. [1] Далее приведен пример сравнения автоматически полученной и авторской аннотаций.

Автоматически полученная аннотация

В настоящей работе для решения краевой задачи для уравнения Пуассона используются различные графические ускорители и библиотека cuFFT для быстрого преобразования Фурье. В настоящей работе решение системы находится с использованием библиотеки CULA, реализующей ряд процедур пакета LAPACK на основе технологии NVIDIA CUDA. Расчеты осуществлялись на системах с различными GPU, в том числе на суперкомпьютере Ломоносов НИВЦ МГУ. Особенность использования этой библиотеки в том, что она не содержит процедур для синус-преобразований, которые в данном случае необходимы для того, чтобы полученное решение удовлетворяло нулевым граничным условиям. Процедура решения будет состоять из следующих этапов. И такие задачи могут возникать при обработке больших БД экспериментальных измерений. В зависимости от ускорителя она составляет от 44 до 150 Гфлопс, что соответствует 410 % от пиковой. Задача моделирования работы масс-спектрометров на основе ионного циклотронного резонанса и преобразования Фурье может быть решена на гибридных системах. В статье представлены результаты реализации кода Pic3D частиц в ячейке для моделирования работы масс-спектрометров на основе ионного циклотронного резонанса и преобразования Фурье на гибридных системах с CPU и GPU. Вычисления показывают, что ускорители могут быть использованы для определения кулоновского взаимодействия ионов с помощью решения первой краевой задачи для уравнения Пуассона и параллельного вычисления полей от каждой поверхности электрода через решение алгебраических систем с достаточной эффективностью.

Авторская аннотация

В работе представлено исследование эффективности параллельных программ для расчета эволюции ионов в рамках модели частиц в ячейке. Разработаны параллельные программы для гибридных вычислительных систем, содержащих устройства CPU (Central Processing Units, центральные процессоры) и GPU (Graphic Processing Units, графические ускорители). Программы применяются для прямого моделирования поведения ионов в ловушках масс-спектрометров на основе преобразования Фурье. Показана возможность использования GPU-устройств для ускорения многократного решения краевых задач для уравнения Пуассона на основе быстрого преобразования Фурье, реализованного в библиотеке cuFFT – библиотеке процедур быстрого преобразования Фурье для архитектуры CUDA (Compute Unified Device Architecture). Проведено сравнение реально достигаемой производительности и задействования полосы пропускания памяти при вычислении решения с пиковыми характеристиками GPU для разных установок. Показано, что выбранный алгоритм решения первой краевой задачи для уравнения Пуассона масштаби-

[1] Международный журнал «Программные продукты и системы». URL: http://www.swsys.ru/.

руется в соответствии с асимптотической оценкой сложности. Разработаны программы расчета полей, удерживающих ионы в ловушке в произвольной геометрии электродов для работы на гибридных системах, сочетающих в себе одновременную обработку данных на CPU и GPU. В каждом из параллельных процессов программы расчета поля решение алгебраических уравнений осуществляется на GPU через процедуры LAPACK, реализованные в составе библиотеки CULA. Результаты расчетов на суперкомпьютере «Ломоносов» показали, что эффективность параллельного использования GPU существенно зависит от выбранной схемы распределения процессов параллельной программы. Ускорители могут эффективно использоваться для определения кулоновского взаимодействия ионов с помощью решения первой краевой задачи для уравнения Пуассона. Параллельные вычисления полей на CPU от каждой поверхности электрода можно проводить совместно с решением алгебраических систем на GPU с достаточной эффективностью.

Пока не существует общепринятого эффективного способа автоматической оценки систем автореферирования [18]. Во-первых, мы пробовали оценить качество полученных аннотаций при помощи метрики ROUGE, основанной на подсчете количества совпадающих элементов в сравниваемых текстах, например, *n*-грамм или предложений [19]. В метрике ROUGE в случае подсчета совпадающих предложений текст аннотации рассматривается как последовательность предложений. Основная идея состоит в том, что чем длиннее самая длинная общая подпоследовательность *LCS* двух предложений в сравниваемых аннотациях, тем более похожими считаются эти две аннотации. Как правило, используют *F*-меру на основе *LCS* для оценки сходства между двумя величинами *X* длиной *m* и *Y* длиной *n*, считая, что *X* является образцом для сравнения, а *Y* – просматриваемый элемент. Точность, полнота и *F*-мера согласно ROUGE определяются следующим образом:

$$P_{lcs} = \frac{LCS(X,Y)}{n}, \quad R_{lcs} = \frac{LCS(X,Y)}{m},$$

$$F_{lcs} = \frac{\left(1+\beta^2\right)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}},$$

где $LCS(X, Y)$ – длина самой длинной общей подпоследовательности X и Y, а $\beta = P_{lcs} / R_{lcs}$.

Были получены следующие значения метрики ROUGE: точность 32,8 %, полнота 59,04 %, *F*-мера 34,47 %. К сожалению, в работах [8; 9], которые описывают системы обработки текстов на русском языке, не приводятся значения метрики ROUGE, поэтому нет возможности сравнить эти результаты с нашими. Также мы пришли к выводу, что некорректно сравнивать результаты работы нашей системы с результатами работы систем для английского языка, такими как, например, [20], поскольку низкие значения ROUGE могут быть связаны с особенностями языкового строя. В частности, русский язык является флективным языком с развитой морфологией, к тому же порядок слов в русском языке относительно свободный.

Во-вторых, мы воспользовались экспертной оценкой. Точность полученных аннотаций, оцененная экспертами, оказалась значительно выше. Экспертная оценка результатов реферирования показала, что 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них (что на самом деле не всегда свидетельствует о плохом качестве реферата), и только 13,57 % представляли собой некорректно отобранные фрагменты текстов. Следует заметить, что полученная нами экспертная оценка выше, чем в работах [8; 9].

Нами было замечено, что авторы часто используют синонимы, перефразируют и меняют местами предложения. Экспертная оценка подтверждает, что порядок предложений в аннотации часто не влияет на ее общий смысл. Однако метрика ROUGE не учитывает это. Кроме того, иногда автоматически сформированная аннотация получается длиннее, чем хотелось бы (около 500 слов вместо 250). Это связано со стилем изложения самой статьи, и чаще всего означает, что в тексте имеется много содержательных предложений.

В-третьих, мы рассмотрели точность, полноту и *F*-меру, которые вычисляются способом, похожим на предложенный в работах [2; 8]. Поясним подробнее. Предположим, что автоматически полученная аннотация содержит в себе множество ключевых слов и многословных терминов W_1, множество специальных слов из научных и технических текстов V_1, множество

дискурсивных маркеров и коннекторов D_1. Объединение этих множеств обозначим N_1: $N_1 = W_1 \cup V_1 \cup D_1$. Аналогичные множества можно выделить в эталонной авторской аннотации N_2: $N_2 = W_2 \cup V_2 \cup D_2$. Тогда точность, полноту и *F*-меру будем вычислять по следующим формулам:

$$P = \frac{|N_1 \cap N_2|}{|N_1|}, \quad R = \frac{|N_1 \cap N_2|}{|N_2|},$$

$$F\text{-measure} = \frac{2 \cdot P \cdot R}{P + R}.$$

Сравнительная оценка результатов приведена в табл. 5.

Таблица 5

Оценка результатов, %

Система	Точность	Полнота	*F*-мера
Scientific Text Summarizer	**75,23**	**68,21**	**71,55**
Trevgoda 2009	67,03	64,81	66,03
Marcu 1998	73,53	67,57	70,42

Преимущество предложенных формул состоит в том, что они позволяют определить вклад каждого из признаков и разных комбинаций этих признаков в общую оценку результата. Например, можно оценить вклад только маркеров и коннекторов или только специальной научной лексики, или того и другого, но без ключевых слов и выражений и т. д. В дальнейшем мы планируем провести подобное исследование данного вопроса.

Возможное улучшение предложенного в данной статье алгоритма, по нашему мнению, состоит в том, чтобы дополнить правила удаления менее важных предложений, увеличить количество шаблонов для сглаживания, расширить списки маркеров, коннекторов и индикаторов.

Заключение

Описан подход к автоматическому построению аннотаций научно-технических текстов на русском языке. Процесс состоит из пяти шагов: предобработка, преобразование текста, оценка весов, выбор предложений и сглаживание. Преобразование предложений включает риторический анализ. На основе дискурсивных маркеров и коннекторов извлекаются наиболее значимые предложения в тексте. Также учитываются ключевые слова, многословные выражения и специальная лексика, которая часто присутствует в научных и технических текстах. Далее из полученного набора предложений для аннотации выбираются только те, вес которых превышает заранее заданную пороговую величину. Операция сглаживания позволяет получить более связный текст. В дальнейшем планируется провести эксперименты с текстами из различных научных областей на других языках.

Список литературы

1. *Ананьева М. И., Кобозева М. В.* Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». М., 2016. URL: www.dialog-21.ru/media/3460/ananyeva.pdf

2. *Marcu D.* Improving summarization through rhetorical parsing tuning // VI Workshop on Very Large Corpora. 1998. P. 206–215.

3. *Hovy E., Lin Ch.-Y.* Automated text summarization and the SUMMARIST system // Proc. of the TIPSTER Text Program. 1998. P. 197–214.

4. *Teufel S., Moens M.* Summarizing scientific articles: experiments with relevance and rhetorical status // Computational Linguistics. 2002. Vol. 28 (4). P. 409–445.

5. *Bosma W.* Query-Based Summarization using Rhetorical Structure Theory // 15th Meeting of CLIN. 2005. P. 29–44.

6. *Huspi S. H.* Improving Single Document Summarization in a Multi-Document Environment. PhD Thesis. Melbourne, Australia: RMIT University, 2017. 190 p.

7. *Mithun S.* Exploiting rhetorical relations in blog summarization. PhD Thesis. Montreal, Canada: Concordia University, 2012. 230 p.

8. *Тревгода С. А.* Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: Дис. … канд. техн. наук. СПб., 2009. 157 с.

9. *Осминин П. Г.* Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод: Дис. … канд. филол. наук. Челябинск, 2016. 239 с.

10. *Batura T. V., Bakiyeva A. M., Yerimbetova A. S. Mit'kovskaya M. V. Semenova N. A.* Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2016. Is. 40. P. 37–51.

11. *Бакиева А. М., Батура Т. В.* Исследование применимости теории риторических структур для автоматической обработки научно-технических текстов // Cloud of Science. Вып. 4, № 3. С. 450–464.

12. *Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. 2017. Iss. 16 (23). Vol. 1. P. 194–204.

13. *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // International Conference on Analysis of Images, Social Networks and Texts (AIST). Ekaterinburg, Russia, 2015. P. 370–384.

14. *Mann W., Thompson C.* Rhetorical structure theory: Toward a functional theory of text organization // Text-Interdisciplinary Journal for the Study of Discourse. 1988. Vol. 8. No. 3. P. 243–281.

15. *Blei D. M., Lafferty J. D.* Visualizing Topics with Multi-Word Expressions // Semantic Scholar. 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

16. *Батура Т. В., Стрекалова С. Е.* Подход к построению расширенных тематических моделей текстов на русском языке // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 5–18.

17. *Vorontsov K.* Welcome to BigARTM's documentation! 2015. URL: http://bigartm.readthedocs.io/en/stable/

18. *Das D., Martins A. A.* Survey on Automatic Text Summarization // Literature Survey for the Language and Statistics II course at CMU. 2007. P. 192–195.

19. *Lin Ch. Y.* ROUGE: A Package for Automatic Evaluation of Summaries // Workshop On Text Summarization Branches Out. 2004. P. 74–81.

20. *Zhang J. J., Chan H. Y., Fung P.* Improving lecture speech summarization using rhetorical information // IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). 2007. P. 195–200.

Материал поступил в редколлегию 02.07.2018

T. V. Batura [1, 2], **A. M. Bakiyeva** [1]

[1] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

[2] *A. P. Ershov Institute of Informatics Systems SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

tatiana.v.batura@gmail.com, m_aigerim0707@mail.ru

DEVELOPING THE SYSTEM FOR AUTOMATIC SUMMARIZATION OF SCIENTIFIC TEXTS

The paper describes a new method of automatic text summarization. Based on this method, a system has been created that makes it possible to obtain summaries of scientific and technical texts and to determine their topics. The summarization process consists of five main steps: preprocessing, transformation, weight evaluation, sentence selection, and smoothing. The proposed method allows receiving the summary based on important sentences of the original document. The importance of sentences is partially determined in the process of rhetorical analysis, which is performed using discursive markers and connectors. Keywords, multiword terms, and some special words that are often found in scientific and technical texts are also taken into account. We used additive regularization for topic modeling (ARTM) to extract keywords and discover the topics.

Keywords: automatic summarization, rhetorical structure theory, discourse markers, additive regularization, topic modeling.

References

1. Ananyeva M. I., Kobozeva M. V. Development the corpus of Russian texts with markup based on the Rhetorical Structure Theory. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"*. Moscow, 2016. URL: www.dialog-21.ru/media/3460/ananyeva.pdf/ (in Russ.)

2. Marcu D. Improving summarization through rhetorical parsing tuning. *VI Workshop on Very Large Corpora*, 1998, p. 206–215.

3. Hovy E., Lin Ch.-Y. Automated text summarization and the SUMMARIST system. *Proc. of the TIPSTER Text Program*, 1998, p. 197–214.

4. Teufel S., Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 2002, vol. 28 (4), p. 409–445.

5. Bosma W. Query-Based Summarization using Rhetorical Structure Theory. *15th Meeting of CLIN*, 2005, p. 29–44.

6. Huspi S. H. Improving Single Document Summarization in a Multi-Document Environment. PhD Thesis. Melbourne, Australia, RMIT University, 2017, 190 p.

7. Mithun S. Exploiting rhetorical relations in blog summarization. PhD Thesis. Montreal, Canada, Concordia University, 2012, 230 p.

8. Trevgoda S. A. Methods and algorithms of automatic text summarization based on the analysis of functional relations. PhD Thesis. St. Petersburg, Russia, 2009, 157 p. (in Russ.)

9. Osminin P. G. Construction of a model for abstracting and annotating scientific and technical texts focused on automatic translation. PhD Thesis. Chelyabinsk, Russia, 2016, 239 p. (in Russ.)

10. Batura T. V., Bakiyeva A. M., Yerimbetova A. S. Mit'kovskaya M. V. Semenova N. A. Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory. *Bulletin of the Novosibirsk Computing Center. Series: Computer Science*, 2016, is. 40, p. 37–51.

11. Bakiyeva A. M., Batura T. V. Research of applicability of the rhetorical structure theory for automatic processing of scientific and technical texts. *Cloud of Science*, 2017, vol. 4, no. 3, p. 450–464. (in Russ.)

12. Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. Towards building a discourse-annotated corpus of Russian. *Computational Linguistics and Intellectual Technologies*, 2017, iss. 16 (23), vol. 1, p. 194–204.

13. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. *International Conference on Analysis of Images, Social Networks and Texts (AIST)*. Ekaterinburg, Russia, 2015, p. 370–384.

14. Mann W., Thompson C. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 1988, vol. 8, no. 3, p. 243–281.

15. Blei D. M., Lafferty J. D. Visualizing Topics with Multi-Word Expressions. *Semantic Scholar*, 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

16. Batura T. V., Strekalova S. E. An Approach to Building Extended Topic Models of Russian Texts. *Vestnik NSU. Series: Information Technologies*, vol. 16, no. 2, p. 5–18. (in Russ.)

17. Vorontsov K. Welcome to BigARTM's documentation! 2015. URL: http://bigartm.readthedocs.io/en/stable/

18. Das D., Martins A. A. Survey on Automatic Text Summarization. *Literature Survey for the Language and Statistics II course at CMU*, 2007, p. 192–195.

19. Lin Ch. Y. ROUGE: A Package for Automatic Evaluation of Summaries. *Workshop On Text Summarization Branches Out*, 2004, p. 74–81.

20. Zhang J. J., Chan H. Y., Fung P. Improving lecture speech summarization using rhetorical information. *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, p. 195–200.

For citation:

Batura T. V., Bakiyeva A. M. Developing the System for Automatic Summarization of Scientific Texts. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 74–86. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-74-86