УДК 004.048:519.765

DOI: 10.15827/0236-235X.129.144-153

Дата подачи статьи: 02.07.19

2020. Т. 33. № 1. С. 144–153

Гибридный метод автореферирования научно-технических текстов на основе риторического анализа

Т.В. Батура^{1,2}, к.ф.-м.н., доцент, старший научный сотрудник,
tatiana.v.batura@gmail.com

А.М. Бакиева³, соискатель, m_aigerim0707@mail.ru

¹ Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск, 630090, Россия

² Новосибирский государственный университет, г. Новосибирск, 630090, Россия

³ Евразийский национальный университет им. Л.Н. Гумилева,
г. Астана, 010008, Республика Казахстан

Непрерывное увеличение интенсивности потока текстовой информации делает все более важной задачу семантического сжатия текстов. Связи между дискурсивными маркерами в тексте задают иерархию, которая позволяет решать различные задачи обработки текстов на естественном языке, в том числе и задачу автоматического составления краткого реферата по тексту большого объема.

В данной статье описан оригинальный метод автореферирования научно-технических текстов на основе риторического анализа и с использованием методов тематического моделирования. Предложенный метод сочетает в себе использование лингвистической базы знаний, графовое представление текстов и машинное обучение.

Для обнаружения ключевых терминов используется тематическое моделирование. Сначала строятся униграммные тематические модели, содержащие только однословные термины. Далее эти модели расширяются путем добавления многословных терминов. Наиболее значимые фрагменты исходного документа определяются в процессе риторического анализа при помощи дискурсивных маркеров. Представление текстов в виде графов помогает наглядно продемонстрировать преобразования, необходимые для выделения важных фрагментов текстов. При оценке важности фрагментов текста также учитываются ключевые слова, многословные термины и научная лексика, характеризующая научно-технические тексты. Для хранения информации о маркерах была создана лингвистическая база знаний. Заключительным этапом формирования аннотации является сглаживание – процедура преобразования текста, позволяющая сделать текст полученного реферата (аннотации) более связным и последовательным.

Результаты экспериментов показали эффективность предложенного метода и возможность его адаптации для обработки текстов разных тематик на других языках.

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

В современном мире наблюдается колоссальный рост количества информации на естественном языке. Разработка алгоритмов и систем автоматического реферирования, поиска и извлечения информации, классификации и кластеризации текстовых документов по-прежнему является сложной задачей.

Непрерывное увеличение интенсивности потока текстовой информации делает все более важной задачу семантического сжатия текстов. Существует много путей ее решения, которые довольно четко подразделяются на три направления: экстракция, абстракция и гибридный подход. Экстракция – извлечение из исходного текста наиболее информативных предложений, то есть формирование квазиреферата. Этот

способ иногда называют поверхностным. К достоинствам экстрагирующих методов можно отнести независимость от предметной области, а также сравнительную простоту разработки: не требуются создание обширных баз знаний и проведение детального лингвистического анализа текста. К недостаткам экстрагирующих методов можно отнести то, что полученные рефераты часто являются бессвязными. Абстракция – генерация текста реферата с учетом морфологии, синтаксиса, семантики, благодаря чему формируется логически и по смыслу связный текст. Этот способ называют глубинным. Преимуществом абстрагирующих методов является получение реферата более высокого качества, чем при применении экстрагирующих

методов. К недостаткам относятся сложность практической реализации методов и необходимость сбора большого количества лингвистических знаний.

Для преодоления недостатков абстрагирующих и экстрагирующих методов разрабатываются гибридные методы автоматического реферирования, сочетающие в себе вышеуказанные подходы. Например, сначала происходят извлечение наиболее значимых фрагментов и их последующая обработка, потом слияние предложений, удаление неинформативных частей и т.д. Сложность при разработке гибридных методов заключается в выборе наиболее удачного сочетания методик генерации и извлечения. Гибридные методы по сравнению с абстрагирующими проще в разработке, а по сравнению с чисто экстрагирующими методами могут обеспечить лучшее качество конечного результата.

Например, в системе COMPENDIUM [1] гибридный подход реализуется следующим образом. На вход подается реферат, составленный по экстрагирующей методике. Для этого реферата строится взвешенный граф, вершины которого представлены словами, а дуги отражают отношения смежности между словами. Вес дуг определяется по алгоритму PageRank. Затем между вершинами графа строится кратчайший путь с помощью алгоритма Дейкстры, таким образом, создается набор предложений-кандидатов. Следующий этап заключается в фильтрации неправильных путей. Авторы выделили критерии правильных предложений: длина предложения не менее трех слов, в каждом предложении должен быть глагол, предложение не должно оканчиваться на артикль, предлог, местоимение или союз. На последнем этапе происходит выбор предложений для включения в новый реферат из реферата, составленного по экстрагирующей методике, или из набора предложений-кандидатов.

Наглядным примером гибридного способа построения системы автореферирования является многоязычная система SUMMARIST, описанная в [2]. Эта система сочетает в себе методы понятийного уровня знаний о мире, методы информационного поиска и статистические методы. Алгоритм состоит из трех этапов: идентификация темы, интерпретация и генерация. SUMMARIST формирует аннотации на пяти языках: английском, японском, испанском, индонезийском и арабском.

Также существует гибридная система SumUM [3], которая генерирует рефераты для

научно-технических документов. Авторы провели исследование корпуса рефератов, выполненных людьми, и выявили ряд трансформаций, которые применяли референты, например, слияние информации из различных частей документа, перефразирование оригинала.

Подход авторов [4] к реферированию основан на поверхностном анализе исходного документа, извлечении информации определенного вида и выполнении генерации текста. В системе также используются маркировщик частей речи (лингвистические и концептуальные шаблоны, заданные регулярными выражениями), синтаксические категории, концептуальный словарь.

В работе [5] предложен метод реферирования, основанный на преобразовании текста в концепты с последующим представлением документа в виде графа. Метод использует дополнительные ресурсы – англоязычный тезаурус медико-биологической области UMLS [6] и программу MetaMap [7] для преобразования текста в концепты из этого тезауруса. Метод заключается в выполнении следующих шагов: представление документа в виде графа, кластеризация концептов, выбор предложений. Прежде всего документ представляется в виде графа, где узлы являются концептами тезауруса UMLS, а ребра обозначают отношения между узлами. Для этого все предложения документа обрабатываются программой MetaMap, концепты UMLS дополняются своими гиперонимами. Далее каждому узлу присваивается оценка, прямо пропорциональная глубине иерархии концептов. После этого все графы предложений объединяются в один граф документа. Затем выполняется кластеризация концептов. Каждый кластер представляет собой набор близких по значению концептов и может рассматриваться как тема документа. Процедура выбора предложений основана на сходстве между кластерами и предложениями. Для выбора предложений авторы используют несколько эвристик.

Естественный язык очень сложен для автоматической обработки, поэтому для улучшения качества получаемых результатов исследователи, как правило, стремятся решать задачи реферирования для определенных предметных областей. Авторы работы [8] исследуют задачу реферирования для текстов судебных решений. Реферированию юридических текстов посвящены, например, работы [9, 10]. Авторы [11] предлагают подход к реферированию оценочных суждений или комментариев пользовате-

лей Интернета. Они собрали корпус оценочных комментариев пользователей из отзывов на сайтах Amazon.com, WhatCar.com и социальной сети Twitter, работали с английским языком, тексты отзывов были посвящены сотовым телефонам и автомобилям. Собранный корпус был вручную размечен экспертом, который определял тональность комментария (отрицательный, нейтральный, положительный) и интенсивность оценки.

Авторы работы [12] предлагают гибридный подход к реферированию текстов патентов на английском, французском и немецком языках. Предложенный в данной работе метод является гибридным. За основу взят дискурсивный анализ текста. Все эксперименты проводились с научно-техническими текстами на русском языке.

Дискурсивный анализ используется довольно широко для решения различных задач компьютерной лингвистики. Подробный обзор литературы, представленный в работе [13], показывает, что в большинстве случаев дискурсивный анализ способен улучшить качество автоматических систем на 4–44 % в зависимости от конкретной задачи.

Система автореферирования научных статей, опирающаяся на дискурсивный анализ, описана в [14]. В ней определены семь риторических категорий. Автор работы [15] 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 [16] обсуждается создание реферата, содержащего не только информацию из одного конкретного документа, но и дополнительные знания из других документов, схожих по тематике.

Как видно, исследования в этой области для английского языка достигли достаточно высокого уровня, но для текстов на русском языке данная область изучена сравнительно мало. Авторам статьи известны лишь диссертационные исследования российских ученых. Так, Тревгода С.А. разработал систему, основанную на правилах вывода и узкоспециализированном словаре ключевых фраз. Гибридный подход, предложенный Осмининим П.Г., реализован в системе, ориентированной на автоматический перевод. Система построена для текстов по теме «математическое моделирование».

В данной статье описана система, реализующая предложенный авторами метод авторе-

ферирования, приведены результаты проводимых экспериментов и дана оценка качества работы.

Риторический анализ и описание преобразований текста

Теория риторических структур – одна из наиболее широко используемых теорий организации текстов [17]. Согласно ей, изначально текст делится на неперекрывающиеся фрагменты, а именно на *элементарные дискурсивные единицы* (ЭДЕ). Последовательные ЭДЕ связаны между собой риторическими отношениями.

В теории риторических структур определяются два типа ЭДЕ: ядро, считающееся наиболее важной частью высказывания, и сателлит, поясняющий ядро и считающийся вторичным. Ядро содержит основную информацию, сателлит – дополнительную информацию о ядре. Сателлит часто непонятен без ядра, а выражения, в которых сателлит удален, могут быть понятны лишь в некоторой степени.

Согласно данной теории, любой текст может быть представлен в виде графа $G = \langle V, E \rangle$, узлами V которого являются элементарные дискурсивные единицы, а ребрами E – отношения между ними. При этом независимо от уровня иерархии узлы графа будут связаны определенным набором отношений как на уровне предложений, так и выше отдельного предложения. Роль таких связей играют риторические отношения (рис. 1).

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, чтобы получился сокращенный текст, являющийся промежуточным между исходным текстом и готовой аннотацией. Для определения границ ЭДЕ используются дискурсивные маркеры. Маркеры (дискурсивные маркеры) – это слова или фразы, которые не имеют реального лексического значения, но зато обладают важной функцией формирования разговорной структуры, передавая намерения говорящего. В зависимости от разных маркеров преобразования текста будут разными, например, удаление предыдущего фрагмента или удаление последующего фрагмента вместе с маркером.

На рисунках 2 и 3 показан фрагмент текста до преобразования и после него.

Общее описание системы

Пусть входной текст состоит из набора предложений $T = [s_1, \dots, s_p]$.

Задача реферирования в том, чтобы найти такое преобразование текста T в реферат $\tilde{T}$, что $\Psi: T \rightarrow \tilde{T}, |T| > |\tilde{T}| \approx 250$ слов. Тогда алгоритм построения реферата можно записать в виде последовательных этапов.

1. Предобработка текста. На этапе предварительной обработки из исходного текста

Рис. 2. Риторический анализ и преобразования графов (до преобразования)

Fig. 2. The rhetorical analysis and graphs mappings (before mapping)

Рис. 3. Риторический анализ и преобразования графов (после преобразования)

Fig. 3. The rhetorical analysis and graphs mappings (after mapping)

удаляются все изображения, таблицы, формулы, информация об авторах и библиографические ссылки.

2. Построение тематических моделей, извлечение ключевых слов и многословных терминов. Тематическое моделирование заключается в построении модели некоторой коллекции текстовых документов. Иначе говоря, тема – это набор слов, тематическая модель – набор тем, упорядоченных по степени вероятности. О них заранее ничего неизвестно, они определяются в результате работы алгоритма (рис. 4).

Униграммная тематическая модель – модель, в которой темы описаны однословными терминами. Многословное выражение (многословный термин) – выражение, состоящее как минимум из двух слов и являющееся синтаксически или семантически уникальным по своей природе. Расширенная тематическая модель – модель, в которой темы описаны не только однословными, но и многословными терминами. Схематично эти понятия представлены на рисунке 5. Для определения тем текстов авторы применяли метод аддитивной регуляризации

Рис. 4. Построение тематической модели коллекции документов

Fig. 4. The topic model building of the document collection

Рис. 5. Униграммная и расширенная тематические модели

Fig. 5. The unigram and extended topic models

тематических моделей ARTM (Additive Regularization for Topic Modeling [18]). Причины выбора и результаты сравнения его с другими методами описаны в работе [19]. Для извлечения многословных терминов был использован алгоритм RAKE (Rapid Automatic Keyword Extraction [20]).

3. Риторический анализ и преобразование текста. На этом шаге обнаруживаются предложения, содержащие дискурсивные маркеры, и выполняются определенные преобразования текста, в результате чего формируется квазиреферат: $T' = [s'_1, \dots, s'_P]$, $T' \subset T$.

В действительности квазиреферат состоит не из предложений в обычном понимании, а из некоторых фрагментов, представляющих собой ЭДЕ. Однако для упрощения изложения в дальнейшем, когда речь идет о квазиреферате, авторы используют термин «предложение».

4. Оценка весов предложений. При вычислении веса каждого предложения квазиреферата учитывается наличие в этом предложении ключевых слов (или многословных терминов), дискурсивных маркеров, а также некоторых слов, характерных для научных текстов. В итоге вес каждого предложения s' вычисляется по следующей формуле:

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

где $W = \{w_1, \dots, w_L\}$ – веса ключевых слов и многословных выражений ($|W| = L$) (веса w_i вычисляются как частоты ключевых слов (или многословных терминов) в тексте); $V = \{v_1, \dots, v_M\}$ – веса значимых глаголов и существительных, часто встречающихся в научных текстах ($|V| = M$) (веса v_j определяются из лингвистической базы знаний); $D = \{d_1, \dots, d_N\}$ – веса дискурсивных маркеров ($|D| = N$) (веса d_k определяются из лингвистической базы знаний).

5. Выбор наиболее важных предложений. Из полученного набора предложений (см. п. 3) для реферата отбираются только те предложения, вес которых (см. п. 4) превышает заданную пороговую величину β :

$$\tilde{T} = [s' \in T' : SW(s') > \beta], \text{ где } \beta = 0,15 - \text{константа, которая определяется эмпирически и от которой зависит, насколько сильно будет сокращен текст.}$$

6. Сглаживание – процедура преобразования текста, позволяющая получать связный текст из разрозненных фрагментов и при необходимости дополнительно сокращать его.

6. Сглаживание – процедура преобразования текста, позволяющая получать связный текст из разрозненных фрагментов и при необходимости дополнительно сокращать его.

Например, в процессе сглаживания заменяются или удаляются некоторые слова или словосочетания, удаляются предложения, длина которых меньше 5 слов, и т.д.

В ходе данной работы была разработана система Scientific Text Summarizer. На рисунке 6 представлена ее блок-схема.

Оценка результатов

Экспертная и автоматическая оценка результатов автореферирования выполнялась на коллекции объемом 1 200 научных статей на русском языке.

Экспертная оценка показала высокие результаты. Эксперт – это человек, который оценивает соответствие содержания исходной статьи тексту автоматически полученной аннотации. 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них, 13,57 % представляли собой некорректно отобранные фрагменты текстов. Считается, что экспертная

оценка зависит от конкретного эксперта, а значит, является субъективной, поэтому наряду с экспертной проводилась автоматическая оценка.

Автоматическая оценка. Качество рефератов оценивалось при помощи точности, полноты и F-меры, введенных аналогично работе [21]. Они вычислялись по следующим формулам:

$$Precision = \frac{|N_1 \cap N_2|}{|N_1|}, Recall = \frac{|N_1 \cap N_2|}{|N_2|},$$

$$F_{measure} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

Здесь N_1 – все ключевые слова, многословные термины, значимая лексика из научных текстов, маркеры, содержащиеся в автоматически полученной аннотации; N_2 – аналогичное множество для авторской аннотации.

Результаты автоматической оценки качества автореферирования представлены в таблице 1, а оценка скорости работы алгоритма для двух коллекций на RAM 6 Гб, Intel Core i5-4210U 1.7 GHz – в таблице 2.

Рис. 6. Блок-схема

Fig. 6. The block diagram

Таблица 1

Автоматическая оценка качества автореферирования

Table 1

Automatic assessment of auto abstracting quality

Система	Метод	Точность, %	Полнота, %	F-мера, %
Коллекция текстов на русском языке				
Trevgoda (2009)	Шаблоны	67,03	64,81	66,03
Open Text Summarizer (2016)	Статистический	12,00	24,20	38,50
Scientific Text Summarizer (2018)	Комбинированный	75,23	68,21	71,55
Коллекция текстов на английском языке				
Marcu (1998)	Комбинация эвристик	73,53	67,57	70,42

Таблица 2

Оценка скорости работы алгоритма

Table 2

Speed assessment of the algorithm work

Операция	Время	
	Для коллекции из 260 текстов	Для коллекции из 1 200 текстов
Предобработка текста	15 сек.	1 мин.
RAKE	5 сек.	20 сек.
ARTM (построение модели)	2 мин.	5 мин.
ARTM (применение готовой модели)	5 сек.	15 сек.
Автореферирование	10 сек.	30 сек.
ВСЕГО	~ 3 мин.	~ 8 мин.

Заключение

Преимуществами предложенного метода автореферирования являются довольно высокие качество и скорость работы, а также наличие небольшого количества данных для обучения, а недостатком – необходимость ручной настройки лингвистической базы знаний. Следует отметить, что при наличии большого количества

формул, рисунков, графиков метод будет работать хуже. По мнению авторов, улучшить предложенный в данной статье алгоритм можно за счет дополнения правил удаления менее важных предложений, увеличения количества шаблонов для сглаживания, расширения списка маркеров. В дальнейшем планируется провести эксперименты с текстами из различных научных областей на других языках.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-01134.

Литература

1. Lloret E., Roma-Ferri M.T., Palomar M. COMPENDIUM: A text summarization system for generating abstracts of research papers. Data & Knowledge Engineering, 2013, vol. 88, pp. 164–175. DOI: 10.1007/978-3-642-22327-3_2.

2. Hovy E., Lin Ch.-Y. Automated text summarization and the SUMMARIST system. Proc. TIPSTER Text Program, 1998, pp. 197–214.

3. Saggion H., Lapalme G. Generating indicative-informative summaries with SumUM. Computational Linguistics, 2002, vol. 28, no. 4, pp. 497–526.

4. Foster G.F. Statistical Lexical Disambiguation. Master’s thes., 1991, 340 p.

5. Plaza L., Diaz A., Gervas P. Concept-graph based biomedical automatic summarization using ontologies. Proc. 3rd Textgraphs Workshop on Graph-Based Algorithms in Natural Language. Manchester, UK, 2008, pp. 53–56.

6. Unified Medical Language System (UMLS). 2016. URL: <http://www.nlm.nih.gov/research/umls/> (дата обращения: 05.06.2019).

7. Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc. AMIA, 2001, pp. 17–21.
8. Farzindar A., Lapalme G. Legal text summarization by exploration of the thematic structures and argumentative roles. Proc. Text Summarization Branches Out Conf., ACL, Barcelona, Spain, 2004, pp. 27–38.
9. Galgani F., Compton P., Hoffmann A. Combining different summarization techniques for legal text. Proc. Workshop on Innovative Hybrid Approaches, EACL, Avignon, France, 2012, pp. 115–123.
10. Megala S., Kavitha A., Marimuthu A. Feature extraction based legal document summarization. Int. J. of Advance Research in Computer Science and Management Studies, 2014, vol. 2, iss. 12, pp. 346–352.
11. Lloret E., Boldrini E., Vodolazova T., Martínez-Barco P., Muñoz R., Palomar M. A novel concept-level approach for ultra-concise opinion summarization. Expert Systems with Applications, 2015, vol. 42, iss. 20, pp. 7148–7156. DOI: 10.1016/j.eswa.2015.05.026.
12. Brüggmann S., Bouayad-Aghab N., Burga A., Carrascosa S., Ciaramella A., Ciaramella M., Codina-Filba J., Escorsa E., Judea A., Mille S., Müller A., Saggion H., Ziering P., Schütze H., Wanner L. Towards content-oriented patent document processing: Intelligent patent analysis and summarization. World Patent Information, 2015, vol. 40, pp. 30–42. DOI: 10.1016/j.wpi.2014.10.003.
13. Ананьева М.И., Кобозева М.В. Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // Компьютерная лингвистика и интеллектуальные технологии: матер. Междунар. конф. 2016. URL: www.dialog-21.ru/media/3460/ananyeva.pdf (дата обращения: 05.06.2019).
14. Teufel S., Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, 2002, vol. 28, pp. 409–445.
15. Bosma W. Query-based summarization using rhetorical structure theory. Proc. 15th Meeting of CLIN, Netherlands, 2005, pp. 29–44.
16. Huspi S.H. Improving single document summarization in a multi-document environment. RMIT Univ. Publ., Melbourne, Australia, 2017, 190 p.
17. Mann W., Thompson C. Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary J. for the Study of Discourse, 1988, vol. 8, no. 3, pp. 243–281. DOI: 10.1515/text.1.1988.8.3.243.
18. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open source library for regularized multimodal topic modeling of large collections. Proc. Int. Conf. AIST, Ekaterinburg, Russia, 2015, pp. 370–384. DOI: 10.25205/1818-7900-2018-16-2-5-18.
19. Батура Т.В., Стрекалова С.Е. Подход к построению расширенных тематических моделей текстов на русском языке // Вестн. НГУ. 2018. Т. 16. № 2. С. 5–18.
20. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. Text Mining: Applications and Theory. 2010, pp. 3–20. DOI: 10.1002/9780470689646.ch1.
21. Marcu D. Improving summarization through rhetorical parsing tuning. Proc. 6th Workshop on Very Large Corpora, Montreal, Canada, 1998, pp. 206–215.

Software & Systems

DOI: 10.15827/0236-235X.129.144-153

Received 02.07.19

2020, vol. 33, no. 1, pp. 144–153

A hybrid method for automatic summarization of scientific and technical texts based on rhetorical analysis

T.V. Batura^{1,2}, *Ph.D. (Physics and Mathematics), Associate Professor, Senior Researcher, tatiana.v.batura@gmail.com*

A.M. Bakieva³, *Applicant, m_aigerim0707@mail.ru*

¹ *A.P. Ershov Institute of Informatics Systems (IIS), Siberian Branch of the Russian Federation Academy of Sciences, Novosibirsk, 630090, Russian Federation*

² *Novosibirsk State University, Novosibirsk, 630090, Russian Federation*

³ *L.N. Gumilyov Eurasian National University, Astana, 010008, Republic of Kazakhstan*

Abstract. Continuously increasing amounts of textual information makes the problem of semantic text pressure more and more important. Relations between discursive markers in a text set the hierarchy, which allows solving the various text-processing problems in natural language, including the problem of the automatic gist abstracting from high-cube text.

This article describes the original method of auto reviewing of scientific and technical texts based on rhetorical analysis and using topic modeling. The suggested method combines the usage of a linguistic knowledge base, graph representation of texts and machine learning.

There is a topic modeling for the detection of essential terms. First, there is a producing of unigram topic models. They contain only one-word terms. Further, these is an expansion of these models by adding multiword terms. The determination of the most important fragments of the source document are in the process of rhetorical analysis with the help of discursive markers. Representation of texts in the form of graphs helps to demonstrate the transformation with the text necessary to extract important fragments. In assessing the importance of text fragments, keywords, multiword terms, and scientific lexicon characterizing scientific and technical texts are also conceded. The creating of the linguistic knowledge base was for the data storage markers. The final step of developing abstract is smoothing – a procedure that makes the text of the summary more coherent and consistent.

As the experiments showed, the proposed method is effective and can be adapted for processing texts of different subjects in other languages.

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

Acknowledgements. The research was with the financial support of RFFP within the frame of a science project no. № 19-07-01134.

References

1. Lloret E., Roma-Ferri M.T., Palomar M. COMPENDIUM: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*. 2013, vol. 88, pp. 164–175. DOI: 10.1007/978-3-642-22327-3_2.
2. Hovy E., Lin Ch.-Y. Automated text summarization and the SUMMARIST system. *Proc. of the TIPSTER Text Program*. 1998, pp. 197–214.
3. Saggion H., Lapalme G. Generating indicative-informative summaries with SumUM. *Computational Linguistics*. 2002, vol. 28, no. 4, pp. 497–526.
4. Foster G.F. *Statistical Lexical Disambiguation*. Master's thes., 1991, 340 p.
5. Plaza L., Diaz A., Gervas P. Concept-graph based biomedical automatic summarization using ontologies. *Proc. 3rd Textgraphs Workshop on Graph-Based Algorithms in Natural Language*. Manchester, UK, 2008, pp. 53–56.
6. *Unified Medical Language System (UMLS)*. 2016. Available at: <http://www.nlm.nih.gov/research/umls/> (accessed June 05, 2019).
7. Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc. AMIA*. 2001, pp. 17–21.
8. Farzindar A., Lapalme G. Legal text summarization by exploration of the thematic structures and argumentative roles. *Proc. Text Summarization Branches Out Conf., ACL*. Barcelona, Spain, 2004, pp. 27–38.
9. Galgani F., Compton P., Hoffmann A. Combining different summarization techniques for legal text. *Proc. Workshop on Innovative Hybrid Approaches, EACL*. Avignon, France, 2012, pp. 115–123.
10. Megala S., Kavitha A., Marimuthu A. Feature extraction based legal document summarization. *Int. J. of Advance Research in Computer Science and Management Studies*. 2014, vol. 2, iss. 12, pp. 346–352.
11. Lloret E., Boldrini E., Vodolazova T., Martínez-Barco P., Muñoz R., Palomar M. A novel concept-level approach for ultra-concise opinion summarization. *Expert Systems with Applications*. 2015, vol. 42, iss. 20, pp. 7148–7156. DOI: 10.1016/j.eswa.2015.05.026.
12. Brüggmann S., Bouayad-Aghab N., Burga A., Carrascosa S., Ciaramella A., Ciaramella M., Codina-Filba J., Escorsa E., Judea A., Mille S., Müller A., Saggion H., Ziering P., Schütze H., Wanner L. Towards content-oriented patent document processing: Intelligent patent analysis and summarization. *World Patent Information*. 2015, vol. 40, pp. 30–42. DOI: 10.1016/j.wpi.2014.10.003.
13. Ananeva M.I., Kobozeva M.V. Development of the corpus of Russian texts with markup based on the Rhetorical Structure Theory. *Proc. of the Intern. Conf. "Dialog 2016"*. Moscow, Russia, 2016. Available at: www.dialog-21.ru/media/3460/ananyeva.pdf (accessed June 05, 2019) (in Russ.).

14. Teufel S., Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*. 2002, vol. 28, pp. 409–445.
15. Bosma W. Query-based summarization using rhetorical structure theory. *Proc. 15th Meeting of CLIN*. Netherlands, 2005, pp. 29–44.
16. Huspi S.H. Improving single document summarization in a multi-document environment. *RMIT Univ. Publ.* Melbourne, Australia, 2017, 190 p.
17. Mann W., Thompson C. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary J. for the Study of Discourse*. 1988, vol. 8, no. 3, pp. 243–281. DOI: 10.1515/text.1.1988.8.3.243.
18. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open source library for regularized multimodal topic modeling of large collections *Proc. Int. Conf. AIST*. Ekaterinburg, Russia, 2015, pp. 370–384.
19. Batura T.V., Strekalova S.E. An approach to building extended topic models of russian texts. *Vestn. NSU. Series: Information Technologies*. 2018, vol. 16, no. 2, pp. 5–18 (in Russ.). DOI: 10.25205/1818-7900-2018-16-2-5-18.
20. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*. 2010, pp. 3–20. DOI: 10.1002/9780470689646.ch1.
21. Marcu D. Improving summarization through rhetorical parsing tuning. *Proc. 6th Workshop on Very Large Corpora*. Montreal, Canada, 1998, pp. 206–215.

Для цитирования

Батура Т.В., Бакиева А.М. Гибридный метод автореферирования научно-технических текстов на основе риторического анализа // Программные продукты и системы. 2020. Т. 33. № 1. С. 144–153. DOI: 10.15827/0236-235X.129.144-153.

For citation

Batura T.V., Bakiyeva A.M. A hybrid method for automatic summarization of scientific and technical texts based on rhetorical analysis. *Software & Systems*. 2020, vol. 33, no. 1, pp. 144–153 (in Russ.). DOI: 10.15827/0236-235X.129.144-153.