ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

2021. Том 14. Выпуск 5. С. 1688-1694 | 2021. Volume 14. Issue 5. P. 1688-1694

ISSN 1997-2911 (print)

Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Особенности использования ключевых терминов в интеллектуальном реферате и научной статье в процессе сжатия текста (на материале английских текстов предметной области «Судостроение»)

Дубинина Е. Ю.

Аннотация. Статья посвящена проблеме сжатия текстовой информации. Цель исследования заключается в уточнении принципов, лежащих в основе операций сжатия текста. Для этого проводится эксперимент, в ходе которого был создан специализированный корпус интеллектуальных рефератов и научных статей, написанных на английском языке и относящихся к предметной области «Судостроение». Научная новизна работы заключается в том, что впервые были обнаружены устойчивые закономерности употребления ключевых терминов. В результате проведения статистического анализа было установлено, что сжатие текста в интеллектуальных рефератах происходит за счет использования многокомпонентных ключевых терминов, имеющих сложные номинативные конструкции. В научной статье эти термины представлены в «усеченном» виде, то есть как двухэлементные или трехэлементные комбинации.

EN

Specificity of Using Professional Terminology in Scientific Articles and Reviews (by the Material of the English-Language Texts of the Subject Area “Shipbuilding”)

Dubinina E. Y.

Abstract. The article examines a text compression algorithm. The research objective involves clarifying principles of the text compression process. To achieve this research objective, the author conducts an experiment during which a corpus of the English-language articles and reviews of the subject area “Shipbuilding” has been developed. Scientific originality of the study lies in the fact that the researcher for the first time discovers principles of professional vocabulary usage in scientific articles and scientific reviews. A statistical analysis allows concluding that in scientific reviews, the text compression algorithm is based on using complicated multicomponent terms. In scientific articles, two- or three-component terms prevail.

Введение

В современном обществе наблюдается экспоненциальное возрастание объемов информации, поступающей к человеку. Темпы количественного роста документов можно проследить по возрастанию числа научных журналов в мире, поскольку они дают до 70% всей научной информации, используемой специалистами. В работе Т. Н. Домниной и О. А. Хачко указываются следующие данные: «…рост количества научных журналов в XX веке составлял не менее 3,3% ежегодно, а в начале XXI века он повысился до 4%» [7, с. 95]. В последние годы быстрый рост Интернета сделал проблему возрастания объема информации еще более острой. Актуальность исследования заключается в необходимости проведения дальнейших поисков путей сжатия текстовой информации.

Теоретической базой данного исследования являются работы в сфере автоматического реферирования Д. И. Блюменау, Н. И. Гендиной, И. С. Добронравова, Д. Г. Лахути, В. П. Леонова, Е. Б. Федорова [4], В. П. Леонова [9], У. Хан (U. Hahn) [19], Х. П. Лун (H. P. Luhn) [20], Д. Марку (D. Marcu) [21]; работы в области интеллектуального реферирования В. И. Горьковой, Э. А. Борохова [5], А. И. Новикова, Н. Л. Сунцовой [11], Е. Кремминс (E. Cremmins) [16]; теории сжатия текстовой информации Д. И. Блюменау [3].

Научная статья (original research article) | https://doi.org/10.30853/phil210207

Задачи исследования:

1) изучение методов и подходов, используемых в процессе сжатия текста;

2) формирование специализированного корпуса текстов определенной тематики, в который включены интеллектуальные рефераты и научные статьи, составленные англоязычными авторами;

3) экстрагирование ключевых терминов из интеллектуальных рефератов и исследование их использования в английских научных статьях, принадлежащих предметной области «Судостроение».

Для решения поставленных задач в работе используются следующие методы исследования: метод сопоставительного анализа, метод статистического анализа, а также методы количественной обработки массивов лингвистических данных.

Практическая значимость работы заключается в возможности применения результатов исследования в семинарских занятиях по английскому языку при обучении студентов реферированию исходного текста; кроме того, полученные данные могут быть использованы при разработке формализованных методов реферирования текста.

Материалом исследования является специально созданный корпус текстов, состоящий из 85 научных статей и рефератов на английском языке, относящихся к единой предметной области «Судостроение». Научные статьи взяты из материалов конференции “Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering” (материалы Международной конференции по морской механике и арктическому машиностроению).

Данная отрасль промышленности за долгие годы развития выработала свою уникальную терминологическую базу для описания объектов, технических процессов, а также различных видов деятельности человека, которые осуществляются на определенных этапах строительства судов. В частности, в английском языке на становление судостроительной терминологии наложили отпечаток исторические процессы, происходившие не только в европейских странах, но и во всем мире. Следовательно, тексты данной предметной области являются релевантной базой для изучения различных языковых явлений.

Методы и подходы, используемые в процессе сжатия текста

В связи с возрастанием массивов текстовой информации эффективность ее переработки посредством традиционных, то есть интеллектуальных, методов значительно уменьшилась. На сегодня возникает все большая потребность в автоматических методах сжатия текста.

Под сжатием понимается сокращение физического объема текстов при условии сохранения содержащейся в них основной информации. Подобная операция осуществляется в системах автоматического реферирования и аннотирования. История применения таких систем насчитывает уже более 50 лет. В последнее время наметились два направления использования вычислительных технологий для реферирования текстового материала. Первым является собственно реферирование, направленное на формирование самостоятельного реферата. Вторым направлением является использование экстрагирующих методов, направленных на извлечение из исходных документов наиболее важных фрагментов [2].

Для создания реферата первого типа требуется проведение большого объема предварительных исследований. Это связано с тем, что в системах, направленных на формирование самостоятельного реферата, содержание текста рассматривается как набор фактов, которые в нем находятся. Для определения содержания текста необходимы следующие виды декларативных знаний:

– знания о мире, воплощенные в именах семантических классов, с помощью которых описывается содержание текстов; эти знания носят тезаурусный характер;

– прагматические знания о тексте, ориентированные на прагматику потребителя;

– знания о содержании текста (представлены в виде заранее заданного набора фактов, значимых для пользователя; эти факты должны распознаваться системой на основе анализа предыдущих двух видов знаний – то есть знаний о мире и прагматических знаний о тексте) [12].

Поскольку для того, чтобы функционировать должным образом, эти системы требуют большого объема информации относительно предметной области, используемые в них методы известны как *knowledge rich methods* – то есть методы, использующие знания. В настоящее время направление представлено разработкой экспериментальных образцов [10; 18; 23].

Второе направление основано на применении различных формальных критериев, позволяющих выделять наиболее информативные фразы (фрагменты) с целью формирования машинного реферата. Для этого используются статистические, позиционные, синтаксические характеристики текста [2; 13; 17]. Исследования, проводимые при создании таких рефератов, являются менее трудоемкими по сравнению с первым направлением. Самый простой тип информации, который успешно используется при автоматическом реферировании текста – это лексическая информация. Наиболее распространенный способ, применяемый в реферировании, – подсчитать частоту использования каждого слова в документе, а затем на этой основе определить «важность», то есть информативность каждого предложения [14]. В некоторых случаях, чтобы получить более надежные результаты, используются различные преобразования.

В основном рабочий процесс по автоматическому реферированию сводится к выбору предложений – «кандидатов» с дальнейшей селекцией для создания реферата. Разработано довольно большое количество процедур, имеющих целью дальнейшую обработку и редактирование предложений. К этим процедурам относятся:

– упорядочивание отобранных предложений;
– введение ограничений на количество предложений в реферате;
– редукция отобранных предложений;
– придание тексту реферата элементов связности [15, p. 360].

Следует отметить, что хотя на сегодня и существует большое количество подходов и методов к автореферированию документов, задача по формированию качественных машинных рефератов еще не решена. Наибольшая сложность заключается в определении информативности исходного документа. Для компьютера этот процесс сложен в связи с тем, что он не только не может «понять» текст, но и обработать текст аналогично человеку даже на наименее сложных уровнях (например, лексическом и синтаксическом). В данной работе предлагается решить эту проблему путем исследования интеллектуальных рефератов и сопоставлениях их с текстами научных статей.

Процесс формирования специализированного корпуса текстов

Для исследования был сформирован специализированный корпус текстов, состоящий из 85 научных статей и рефератов на английском языке. При его создании были учтены следующие признаки корпуса:
– расположение на машинном носителе;
– наличие единой процедуры отбора лексического материала;
– репрезентативность [8].

Ниже эти признаки рассмотрены более подробно. Все тексты, включенные в корпус, были размещены на машинном носителе, то есть представлены в электронном виде. Это дало возможность обрабатывать корпус, применяя компьютерные технологии. В частности, для определения частоты встречаемости ключевых терминов использовалась программа AntConc [22].

Под единством процедуры отбора материала подразумевалось, что в корпус включались только те тексты, которые соответствовали определенным требованиям, а именно:
– текст должен принадлежать единой предметной области («Судостроение»);
– текст должен быть составлен англоязычными авторами и, соответственно, написан на английском языке;
– текст должен быть фиксированного объема;
– текст должен соответствовать определенной структуре: в нем должны присутствовать такие элементы, как заголовок, реферат и, собственно, сам текст статьи.

Следующим признаком корпуса является репрезентативность, под которой понимается «способность корпуса текстов отражать все свойства проблемной области, релевантные для данного типа лингвистических исследований, в определенной пропорции, определяемой частотой явления в проблемной области» [1, с. 38]. Отметим, что все тексты, входящие в корпус, относятся к единой предметной области. Также они относятся к одному типу текстов – научная статья. Данное единообразие позволяет считать корпус репрезентативным и дает возможность переносить информацию, полученную на основе исследования этого корпуса текстов, на всю предметную область. В результате был создан специализированный корпус, на базе которого исследовался процесс сжатия текста.

Экстрагирование ключевых терминов из интеллектуальных рефератов и исследование их использования в научной статье

В данной работе предполагается, что ключевые элементы, содержащиеся в интеллектуальном реферате, определяют наиболее важную информацию текста и могут использоваться для уточнения методических принципов, лежащих в основе операций сжатия текста. Такими ключевыми элементами могут служить терминологические субстантивные словосочетания. Если выделить такие элементы из реферата и проанализировать их использование в самом тексте, то можно выявить механизм сжатия текста.

Из текстов заголовков и рефератов экстрагировались ключевые термины. При этом была использована специально разработанная методика, описанная ниже. На начальном этапе для каждой научной статьи из корпуса был создан словник всех лексических единиц, используемых в заголовке и реферате. Затем из созданного словника исключались стоп-слова, под которыми подразумевается служебная лексика и слова, которые не относятся к определенной тематике [6].

Далее на базе оставшихся в словнике лексических единиц были выделены субстантивные словосочетания (именные группы), которые принято считать ключевыми терминами. На следующем этапе анализировалось использование ключевых терминов в научной статье. Для этого была вычислена частота встречаемости каждого термина в заголовке, реферате, а также в тексте статьи. Кроме того, определялся количественный состав каждого термина, то есть количество входящих в него элементов. Были выделены ключевые термины, в состав которых входили 2, 3 и более компонентов.

По полученным результатам были сформированы алфавитно-частотные таблицы ключевых терминов. Ниже приведен пример такой таблицы (см. Таблицу 1).

Таблица 1. *Ключевые термины, экстрагированные из текста заголовка и реферата*

Ключевые термины	Частота в компонентах статьи			Общая частота
	Заголовок	Реферат	Статья	
containership (контейнеровоз)	1	1	14	16
fatigue cracks (усталостные трещины)	1	1	0	2
fatigue damage (усталостное повреждение)	1	3	39	43
high-frequency damage (повреждение, вызванное высокочастотной вибрацией)	0	1	11	12
large ocean-going ships (большие океанские суда)	0	1	1	2
nonlinear hydroelastic strip theory (нелинейная теория гидравлического сопротивления)	0	1	1	2
springing and whipping effects (влияние спрингинга и випинга)	0	1	0	1
springing contribution (пружинящая реакция)	0	1	2	3
steady wave (стационарная волна)	0	2	11	13
three-dimensional (3D) effects (трехмерное воздействие)	0	1	0	1
total fatigue damage (общее усталостное повреждение)	0	2	4	6
two-dimensional (2D) slamming calculation (двухмерный расчет слеминга)	0	1	0	1
vibrations (вибрации)	0	1	7	8
wave frequency damage (повреждение, вызванное длиной волны)	0	1	4	5
wave-induced vibrations (волновые вибрации)	1	4	8	13

Также были сформированы таблицы, которые отражали частотные характеристики распределения ключевых терминов в корпусе текстов. Такие таблицы были составлены для ключевых терминов, содержащих 2, 3 и более компонентов (см. Таблицы 2, 3, 4). В данных таблицах указана частота использования ключевых терминов в заголовке/реферате и статье, а также общая частота.

Таблица 2. *Частотные характеристики распределения двухкомпонентных ключевых терминов в корпусе текстов (фрагмент)*

	Частота в компонентах статьи		Общая частота
	Заголовок / реферат	Статья	
1	32	130	162
2	4	36	40
3	19	70	89
4	21	44	65
5	17	57	74
6	8	18	26
7	14	58	72
8	13	23	36
9	11	9	20
10	11	15	26

Таблица 3. *Частотные характеристики распределения трехкомпонентных ключевых терминов в корпусе текстов (фрагмент)*

	Частота в компонентах статьи		Общая частота
	Заголовок / реферат	Статья	
1	7	4	11
2	3	1	4
3	13	32	45
4	4	1	5
5	10	12	22
6	4	1	5
7	3	2	5
8	8	5	13
9	5	0	5
10	5	6	11

Таблица 4. *Частотные характеристики распределения четырехкомпонентных ключевых терминов в корпусе текстов (фрагмент)*

	Частота в компонентах статьи		Общая частота
	Заголовок / реферат	Статья	
1	5	1	6
2	6	2	8
3	2	1	3
4	4	1	5
5	3	1	4

	Частота в компонентах статьи		Общая частота
	Заголовок / реферат	Статья	
6	1	0	1
7	3	1	4
8	4	0	4
9	5	0	5
10	7	3	10

На базе этих таблиц был проведен сравнительный анализ состава ключевых терминов в заголовке, реферате и статье. Было выявлено, что наибольшую долю в рефератах и научных статьях составляют ключевых термины, содержащие два компонента (средняя частота употребления: 14 – в интеллектуальном реферате и 39 – в научной статье). Частота употребления терминов, имеющих в составе три компонента, также является относительно высокой как в реферате, так и в статье (средняя частота употребления: 6 – в интеллектуальном реферате и 15 – в научной статье).

Ключевые термины, имеющие в составе четыре компонента (и более), употребляются авторами, как правило, только в заголовке/реферате, в самой же статье частота их употребления стремится к нулю. В частности, было выявлено, что четырехсоставные ключевые термины используются в 22% статей; пятисоставные – в 11% статей; шестисоставные – в 1% статей; семисоставные – не используются.

Данные многокомпонентные ключевые термины в статье находятся в «усеченном» виде. В Таблице 5 приведены примеры таких конструкций.

Таблица 5. *Многокомпонентные и малокомпонентные ключевые термины в реферате и статье (фрагмент)*

	Многокомпонентные ключевые термины: заголовок/реферат	Малокомпонентные ключевые термины: статья
1	high-speed Air Cushion Vehicle (скоростное судно на воздушной подушке)	Air Cushion Vehicle (судно на воздушной подушке)
2	high speed, hard chine hull form (скоростное остроскулое судно)	hull form (форма корпуса судна)
3	upright and heeled wind heeling moment (вертикальный и ветровой накреняющий момент)	heeling moment (накреняющий момент)
4	advanced Computational Fluid Dynamic (CFD) analysis (расширенный анализ вычислительной гидродинамики)	CFD analysis (анализ вычислительной гидродинамики)
5	single-phase level set free surface (однофазный набор уровней свободной поверхности)	free surface (свободная поверхность)
6	center plane maximum wave elevation (максимальная высота волны в центральной плоскости)	wave elevation (высота волны)

Итак, в результате количественных подсчетов были обнаружены следующие закономерности в использовании ключевых терминов:

– концентрация многокомпонентных ключевых терминов в интеллектуальных рефератах; были выявлены структуры, имеющие в составе 8 компонентов;

– преимущественное присутствие в научных статьях этих же терминов в «усеченном» виде, то есть как двухэлементные или трехэлементные комбинации.

Таким образом, мы можем предположить, что количественный состав ключевых терминов меняется в зависимости от назначения текста. При написании реферата одной из задач автора является размещение большого количества информации в малом объеме, что и достигается за счет использования многокомпонентных ключевых терминов. В статье требуется детально представить и объяснить ход исследования, поэтому для облегчения восприятия текста ключевые термины представлены в «усеченном» виде.

Заключение

В результате проведенного исследования, цель которого состояла в уточнении методических принципов, лежащих в основе операций сжатия текста, были сделаны следующие выводы.

Изучение методов и подходов, используемых при сжатии текста, дает основание полагать, что большая часть исследований относится к созданию рефератов, составленных на основе использования формальной структуры текста. Этот вид реферирования является менее трудозатратным. При этом наибольшая сложность состоит в определении информативности исходного текста.

Созданный корпус английских текстов по предметной области «Судостроение», в состав которого были включены интеллектуальные рефераты и научные статьи, дает возможность получить объективную информацию о механизме сжатия текста. В процессе создания данного корпуса были учтены все признаки, которые позволяют назвать массив текстов корпусом. Исследование в рамках созданного специализированного корпуса позволило установить основные языковые механизмы, которые определяют процесс сжатия текста.

Выбор в качестве ключевых элементов терминологических субстантивных словосочетаний, экстрагированных из интеллектуальных рефератов, отвечает цели исследования. При частотном анализе употребления данных сочетаний в интеллектуальных рефератах и научных статьях были выявлены четкие закономерности. В рефератах сконцентрированы термины, имеющие в составе четыре, пять и более компонентов. Эта особенность объясняется назначением данного типа текста: разместить в малом объеме как можно больше полезной информации. В научных статьях для облегчения восприятия текста читателем эти же термины представлены в «усеченном» виде, то есть как двух- и трехкомпонентные сочетания. При автоматизации процесса реферирования, вероятно, наиболее перспективным является применение именно таких ключевых сочетаний. Выявленные данные могут иметь прикладное значение при разработке формализованных методов реферирования документов.

Перспективы дальнейшего исследования заключаются в лингвостатистическом анализе распределения малокомпонентных ключевых терминов в научной статье на базе данных о структурной разметке корпуса текстов.

Источники | References

1. Баранов А. Н. Введение в прикладную лингвистику. М.: Едиториал УРСС, 2009. 360 с.
2. Батура Т. В., Бакиева А. М. Методы и системы автоматического реферирования текстов. Новосибирск: ИПЦ НГУ, 2019. 110 с.
3. Блюменау Д. И. Информационный анализ/синтез для формирования вторичного потока документов. СПб.: Профессия, 2002. 240 с.
4. Блюменау Д. И., Гендина Н. И., Добронравов И. С., Лахути Д. Г., Леонов В. П., Федоров Е. Б. Формализованное реферирование с использованием словесных клише (маркеров) // Научно-техническая информация. Серия 2. Информационные процессы и системы. 1981. № 2. С. 16-20.
5. Горькова В. И., Борохов Э. А. Реферат в системе научной коммуникации. Направления совершенствования лингвистических и структурных характеристик. М.: ВИНИТИ, 1987. 323 с.
6. Гращенко Л. А. О модельном стоп-словаре // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. 2013. № 1 (150). С. 40-46.
7. Домнина Т. Н., Хачко О. А. Научные журналы: количество, темпы роста // Информационное обеспечение науки: новые технологии: сб. науч. тр. М.: БЕН РАН, 2015. С. 83-96.
8. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб.: Изд-во С.-Петерб. ун-та, 2020. 234 с.
9. Леонов В. П. Реферирование и аннотирование научно-технической литературы. Новосибирск: Наука, 1986. 175 с.
10. Лукашевич Н. В. Представление знаний в системе автоматической обработки текстов // Научно-техническая информация. Серия 2. Информационные процессы и системы. 1997. № 3. С. 27-33.
11. Новиков А. И., Сунцова Н. Л. Концептуальная модель порождения вторичного текста // Обработка текста и когнитивные технологии. 1999. № 3. С. 158-166.
12. Откупщикова М. И., Кремнева Н. Д., Кириченко Н. Л. Функционально-семантическая информация в словарных процедурах для анализа текстов узкой предметной области // Структурная и прикладная лингвистика. 1993. № 4. С. 181-196.
13. Тарасов С. Д. Современные методы автоматического реферирования // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2010. № 6 (113). С. 59-74.
14. Andonov F., Slavova V., Petrov G. On the Open Text Summarizer // Information Content and Processing. 2016. Vol. 3. P. 278-287.
15. Babar S., Pallavi D. Improving Performance of Text Summarization // Procedia Computer Science. 2015. Vol. 46. P. 354-363.
16. Cremmins E. The Art of Abstracting. 2nd ed. Arlington, VA: Information Resources Press, 1994. 230 p.
17. Elhadi M. Extractive Summarization Using Structural Syntax, Term Expansion and Refinement // International Journal of Intelligence Science. 2017. Vol. 7. P. 55-71.
18. Goldstein A., Shahar Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data // Journal of Biomedical Informatics. 2016. Vol. 61. P. 159-175.
19. Hahn U. Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge Based Abstraction // Advances in Automatic Text Summarization. The MIT Press, 1999. P. 215-232.
20. Luhn H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information // IBM Journal of Research and Development. 1957. Vol. 1. P. 309-317.
21. Marcu D. The automatic construction of large-scale corpora for summarization research // Proceedings of the 22nd International Conference on Research and Development in Information Retrieval. 1999. August. P. 137-144.
22. Nation P., Anthony L. Measuring vocabulary size // Handbook of Research in Second Language Teaching and Learning: in 3 vols. / ed. by E. Hinkel. N. Y.: Routledge, 2016. Vol. III. P. 355-368.
23. Sahoo D., Bhoi A., Balabantaray R. Hybrid Approach to Abstractive Summarization // Procedia Computer Science. 2018. Vol. 132. P. 1228-1237.

Информация об авторах | Author information

RU

Дубинина Екатерина Юрьевна[1], к. филол. н.
[1] Санкт-Петербургский государственный университет аэрокосмического приборостроения

EN

Dubinina Ekaterina Yurievna[1], PhD
[1] Saint Petersburg State University of Aerospace Instrumentation

[1] *eka609@yandex.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 23.03.2021; опубликовано (published): 31.05.2021.

Ключевые слова (keywords): автоматическое реферирование; интеллектуальный реферат; ключевые термины; научная статья; сжатие текста; automatic abstracting; scientific review; key terms; scientific article; text compression.