

ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ ТЕКСТА НАУЧНОЙ СТАТЬИ В ПРОЦЕССЕ СОЗДАНИЯ АВТОМАТИЧЕСКОГО РЕФЕРАТА

Е. Ю. Дубинина

*Санкт-Петербургский государственный университет
аэрокосмического приборостроения*

Поступила в редакцию 24 октября 2019 г.

Аннотация: в статье рассматриваются вопросы, возникающие в процессе автоматической обработки текста. В этой связи описывается процесс автоматического создания рефератов, который основан на компрессии исходного текста за счет выбора из него предложений с ключевыми словами. На основе сравнительного исследования корпуса текстов научных статей и рефератов осуществляется выделение наиболее информативных лексических единиц.

Ключевые слова: автоматическая обработка текста, автоматическое реферирование, именные группы, ключевые термины.

Abstract: the paper deals with the questions arising in the process of automatic text processing. Thereupon the process of automatic creation of abstracts is under consideration, the process is based on the compression of the source text due to selecting from it sentences with key words. On the basis of a comparative analysis of scientific articles and author's abstracts, the most informative lexical units are selected.

Keywords: automatic text processing, automatic summarization, noun phrases, key terms.

Появление огромных массивов данных, поступающих к человеку с высокой скоростью, вызывает необходимость создания соответствующих программных продуктов для их переработки. В этой связи важной задачей представляется повышение качества систем, осуществляющих автоматическую обработку текста, в частности, систем автоматического реферирования, которые позволяют создавать рефераты, отражающие наиболее важную текстовую информацию.

На сегодня большинство систем автоматического реферирования базируется на компрессии исходного текста за счет выбора из него предложений с ключевыми словами. В лингвистическом аспекте под ключевыми словами понимаются слова / словосочетания, которые являются носителями наиболее важной, существенной информации, содержащейся в тексте [1], [2].

Необходимо отметить, что ключевые слова имеют большое значение в системах научного обслуживания. В частности, они могут использоваться для идентификации и нахождения документов в информационно-поисковых системах. Кроме того, ключевые лексические единицы (ключевые термины) могут являться средством ориентации читателя в текстовом пространстве. Так, например, используя список ключевых слов к научной статье, читатель может подобрать те тексты, которые входят в сферу его исследования.

Помимо этого, ключевые слова используются в современных системах автореферирования. В боль-

шинстве таких систем реализуется следующая процедура: по заранее заданным признакам выделяются наиболее информативные слова, которые считаются ключевыми. Затем осуществляется выбор текстовых фрагментов (в основном, предложений), в которых присутствуют эти ключевые слова. Далее каждому информативному предложению присваивается определенный вес в зависимости от его длины и наличия в нем ключевых слов. На этой основе в реферат отбираются предложения с наиболее высоким весом [3], [4], [5].

Выделение ключевых слов можно производить и «вручную», но, учитывая постоянно растущие объемы текстовой информации — это очень трудоемкий процесс. Кроме того, сложно избежать некоторой субъективности при оценке значимости ключевых слов, так как стандарты для их выделения еще не разработаны [6], [7].

Сейчас существуют различные автоматические методы выделения и оценки значимости ключевых слов: статистические, позиционные, методы фрейм-ового анализа, синтаксические [8], [9]. Однако, несмотря на существование большого количества методов, задача экстрагирования ключевых слов из текста и оценка их значимости еще не решена.

В данной работе выделение ключевых слов и оценка их значимости производится на базе сравнительного анализа корпуса текстов научных статей и их рефератов. В результате предварительного анализа было отобрано 100 текстов, принадлежащих англоязычным авторам из Австралии, Великобритании, Канады и США. Все тексты относятся к предметной области

«Судостроение» и состоят из следующих элементов: заголовок, реферат и научная статья, в которой выделены вводная, основная и заключительная часть.

Для определения ключевых слов из созданного корпуса текстов заголовков и рефератов выделялись простые именные группы. Для их выделения составлялся словарь антипризнаков, или словарь стоп-слов, в который входила служебная лексика, а также слова с общим и общенаучным значением [10]. Эта лексика при дальнейшем анализе не учитывалась. Затем на основе оставшихся лексических единиц были отобраны простые именные группы, которые считались ключевыми терминами и использовались для дальнейшего исследования.

Эти ключевые термины анализировались относительно их использования авторами в научных

статьях. В ходе исследования были сформированы таблицы. Так, например, составлялись таблицы, в которых указывались следующие данные: частота встречаемости ключевого термина в заголовке, реферате и в компонентах статьи (т.е. во вводной части, основной части и заключительной части). Пример таблицы представлен ниже (см. табл. 1). Подсчитывался также уровень распространенности каждого ключевого термина. Под уровнем распространенности понимается количество вхождений ключевого термина из заголовка и/или реферата в компоненты статьи. Этот показатель использовался для определения значимости ключевых терминов.

Кроме того, была сформирована таблица, в которой указывались количественные характеристики ключевых терминов в массиве текстов.

Таблица 1.

Ключевые термины, выделенные при сопоставлении заголовка, реферата и научной статьи (фрагмент)

Ключевые термины	Частота в компонентах статьи					Общая частота	Распространенность
	Заголовок	Реферат	Вводная часть	Основная часть	Заключительная часть		
boundary integral equation		2	2	1	1	6	4
direct potential method		1	3	4	1	9	4
exciting forces		1	1	3	2	7	4
fast hull forms		1				1	1
hydrodynamic coefficients		1	2	7	2	12	4
line integral terms	1	2	3	17	3	26	4
source distribution method		1	3	13	2	19	4
three-dimensional seakeeping characteristics	1					1	1
water line integral terms		1	1			2	2

Для того чтобы определить значимость выделенных ключевых терминов, в работе был использован такой показатель, как уровень распространенности. Как отмечалось выше, ключевыми терминами считались именные группы, выделенные как из заголовка, так и из реферата, поэтому эти компоненты статьи можно считать одинаково информативными. Исходя из этого, было принято, что, если ключевой термин расположен в заголовке и/или реферате, то уровень его распространенности равен 1. Если ключевой термин встречался помимо заголовка и/или реферата в каком-либо другом компоненте статьи (например, во вводной части), то уровень распространенности принимался равным 2.

Если этот же термин встречался во всех компонентах статьи (то есть в заголовке и/или реферате, вводной, основной и заключительной части), то ему присваивался уровень распространенности равный 4.

Таким образом, максимальный уровень распространенности не может быть больше 4.

Исходя из вышесказанного, составленные таблицы анализировались относительно уровня распространенности ключевых терминов.

Анализ составленных таблиц показал, что ключевые термины, имеющие высокий уровень распространенности, являются в основном двух- и трехкомпонентными именными группами, характеризующими предметную область в целом. Например: *boundary element method*, *hydrodynamic forces*, *model towing resistance*.

Относительно частоты ключевых терминов никаких закономерностей обнаружить не удалось. В авторских рефератах использовались термины и с достаточно высокой частотой встречаемости (более 27), и низкой (равной 3 или 4). Таким образом, частота ключевого термина не является показателем

его информативности, более важным является уровень его распространенности.

Кроме того, выделенные ключевые термины анализировались относительно их распределения по структурным разделам каждой статьи. При этом использовалось положение о рамочной конструкции текста. В соответствии с этим положением признаком информативности элемента текста является его пространственно-позиционная характеристика, то есть расположение в тексте. Исходя из вышеизложенного, было сделано предположение, что в тексте научной статьи высокоинформативными являются следующие элементы: заголовки, а также вводная и заключительная часть текста.

Для проверки этого предположения отдельно рассматривалась основная часть текста. Анализировались ключевые термины, которые встречались одновременно в заглавии, реферате и в основной части текста. Количественный анализ именных групп, выделенных в заголовке, реферате и основной части показал, что лишь незначительная часть ключевых терминов, используемых в рефератах, содержится в основной части текста. Таким образом, учет рамочной конструкции текста позволяет упростить поиск релевантных сведений и сократить объем текстовой информации, подлежащей переработке.

Выделение ключевых слов и оценка их значимости может производиться на базе сравнительного анализа корпуса текстов научных статей и рефератов. При этом для определения степени информативности ключевых слов можно использовать такой показатель, как уровень распространенности.

Кроме того, при оценке значимости ключевых слов необходимо учитывать их расположение в тексте. Исследование особенностей размещения ключевых слов в структурных компонентах научной статьи показывает, что слова, входящие в текст заголовка, вводную и заключительную часть текста, являются высокоинформативными.

Итак, на основании сравнительного исследования корпуса текстов научных статей и рефератов было сделано заключение о наиболее информативных ключевых словах. Наиболее информативными являются следующие ключевые слова:

- ключевые слова, расположенные в заголовке, вводной и заключительной части текста;
- ключевые слова, имеющие высокий уровень распространенности;
- ключевые слова, состоящие из двух или трех компонентов.

Именно эти слова являются ключевыми и на их основе с помощью специальных алгоритмических процедур возможно формирование автоматического реферата.

ЛИТЕРАТУРА

1. Гринева М. П. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов / М. П. Гринева. — Режим доступа: http://citforum.ru/database/articles/kw_extraction/
2. Москвитина Т. Н. Ключевые слова и их функции в научном тексте / Т. Н. Москвитина // Вестник Челябинского государственного педагогического университета. — 2009. — № 11. — С. 270–283.
3. Солошенко А. Н. Автоматизация реферирования новостных Интернет-текстов / А. Н. Солошенко, Ю. А. Орлова, В. Л. Розалиев // Известия Волгоградского государственного технического университета. — 2013. — № 22(125). — С. 81–86.
4. Тарасов С. Д. Метод тематического связанного ранжирования для задач автоматического сводного реферирования научно-технических информационных сообщений: автореф. дис. ... канд. техн. наук: 05.13.01 / Тарасов С. Д.; Балт. гос. техн. ун-т «Военмех» им. Д. Ф. Устинова. — СПб., 2011. — 21 с.
5. Niles R. Automatic Text Summarization with Cohesion Features / R. Niles // International Journal of Computer Science and Information Technologies. — 2017. — Vol. 8, no. 2. — Pp. 194–198.
6. Абрамов Е. Г. Подбор ключевых слов для научной статьи / Е. Г. Абрамов // Научная периодика: проблемы и решения. — 2011. — № 2. — С. 35–40.
7. Камшилова О. Н. Малые формы научного текста: ключевые слова и аннотация (информационный аспект) / О. Н. Камшилова // Известия Российского государственного педагогического университета им. А. И. Герцена. — 2013. — № 156. — С. 106–117.
8. Лукашевич Н. В. Комбинирование признаков для автоматического извлечения терминов / Н. В. Лукашевич, Ю. М. Логачев // Вычислительные методы и программирование: Новые вычислительные технологии. — 2010. — № 11(2). — С. 108–116.
9. Bharti S. Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles / S. Bharti, K. Babu, A. Pradhan // European Journal of Advances in Engineering and Technology. — 2017. — Vol. 4, no. 6. — Pp. 410–427.
10. Беляева Л. Н. Лингвистические автоматы в современных гуманитарных технологиях / Л. Н. Беляева. — СПб.: Книжный дом, 2007. — 192 с.

*Санкт-Петербургский государственный университет
аэрокосмического приборостроения*

*Дубинина Е. Ю., доцент кафедры иностранных языков
E-mail: eka609@yandex.ru*

*Saint Petersburg State University of Aerospace Instrumentation
Dubinina E. Yu., Associate Professor of the Department of
Foreign Languages
E-mail: eka609@yandex.ru*