ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Д. В. Степанова, М. А. Жданович

ОСОБЕННОСТИ ВЫДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ ИЗ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ ПРЕДМЕТНОЙ ОБЛАСТИ «ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»

В данной статье рассматривается специфика применения статистического подхода для выделения ключевых слов из научно-технических текстов. В качестве основного критерия для решения поставленной задачи была принята частотность, определяемая посредством программного продукта для семантико-статистического анализа текста Istio. Приводится подробное описание этапов статистической обработки разноплановых текстов, которая позволила составить список ключевых слов. Их анализ единиц показал, что 93,4 % отобранных слов являются ключевыми для текстов одного источника. Проверка отобранных ключевых слов по толковым словарям подтверждает необходимость обработки текстов различных источников и эффективность применения программного обеспечения для оперативного извлечения ключевых слов.

В современном мире во всех сферах деятельности общества наблюдается интенсификация развития науки и техники, что сопровождается распространением больших объемов информации, содержащей значительное количество новых терминов, не отраженных в словарях. Следовательно, существует постоянная необходимость в пополнении терминологических словарей новыми лексическими единицами.

При проведении лингвистических исследований по выделению терминов из научно-технических текстов применяются различные подходы. В целом для достижения указанной цели принято выделять семантический и статистический подходы [1]. Применение семантического подхода заключается в установлении логико-семантической соотнесенности понятия определенной предметной области и его плана выражения. Статистический подход реализуется посредством применения к текстам специально разработанных формул. В данном случае одним из главных факторов является частотность употребления слов. Сопутствующие факторы зависят от целей исследования и установок автора. По мнению А. В. Зубова, реализация статистических методов позволяет получить более объективный и более достоверный результат в процессе выделения терминов [1].

В рамках проводимого исследования по упорядочению терминологии предметной области «Информационные технологии» (ИТ) для решения задачи по выделению актуальной терминологии целесообразным представлялось использование комбинированного подхода, т.е. сочетания статистического и семантического подходов, реализация которого осуществляется в два этапа. Первый этап предполагает учет статистических характеристик, а именно частоты и частотности словоупотреблений, для

определения ключевых слов в исследуемых текстах. Второй этап заключается в проведении компонентного анализа дефиниций отобранных ключевых слов для определения степени их терминологичности и составления перечня актуальных терминов заданной предметной области. В рамках представленной статьи рассматриваются особенности применения статистического подхода к выделению ключевых слов из научно-технических текстов.

Источниками материла для исследования послужили тексты научных статей англоязычных журналов 2016–2019 гг. издания по различным аспектам предметной области ИТ [2; 3; 4], а также тексты руководств по эксплуатации новейшего программного обеспечения компании Microsoft за 2019 г. [5]. Для дальнейшего анализа отобранного материала были построены исследовательские корпусы текстов по каждому источнику. При составлении корпуса текстов периодических изданий было принято решение отбирать только названия статей, их аннотации и списки ключевых слов, что позволило повысить концентрацию терминов в создаваемом корпусе, снижая вероятность вхождения в частотные словари общеупотребительных слов (общий объем разработанного корпуса составляет 629 524 словоупотреблений). В корпус текстов руководств по эксплуатации компании Microsoft были включены тексты без предварительной обработки методом сплошной выборки (общий объем разработанного корпуса составляет 86 788 словоупотреблений). Полученные корпусы текстов были разделены на 20 подкорпусов, в том числе 12 подкорпусов, которые включают материал 3 журналов за 4 года издания, и 8 подкорпусов, которые выделены в соответствии с описываемым программным обеспечением (Excel, Microsoft Team, OneDrive, OneNote, Outlook, PowerPoint, SharePoint и Word).

Для решения задачи по автоматизированному извлечению ключевых слов из научно-технических текстов по ИТ был выбран программный продукт для статистического анализа текстов Istio, позволяющий получить информацию о частоте и частотности употребления лексических единиц. При этом показатель частотности (I_f) выражает отношение количества употреблений отдельного слова в тексте, или его частоту (F), к общему количеству слов в тексте (N) и высчитывается по следующей формуле:

$$If = \frac{F}{N} \cdot 100\ \%$$

Например, в одном из подкорпусов текстов, который включает 53 113 словоупотреблений, слово *datum* встречается 584 раза (F). Следовательно, частотность (I_f) составляет 1 %. Предварительный анализ текстов показал, что целесообразным представляется фиксация в качестве ключевых тех слов, показатель частотности которых равен или превышает 0,3 %.

Следует отметить, что в англоязычных текстах научного стиля высокой частотностью обладают служебные части речи, которые в рамках статистической обработки текстов относятся к стоп-словам. Например, в указанном ранее подкорпусе наивысший показатель частотности имеет артикль *the*.

В связи с этим было принято решение удалить все стоп-слова из полученного частотного словаря, что также было выполнено автоматически посредством сервиса Istio.

Отобранные для исследования 20 подкорпусов текстов подлежали статистическому анализу с помощью сервиса Istio поэтапно. Вначале были проанализированы 4 подкорпуса текстов журнала «Future Generation Computer Systems» (табл. 1): 2016 г. – 53 113 словоупотреблений, 2017 г. – 65 460 словоупотреблений, 2018 г. – 160 314 словоупотреблений, 2019 г. – 186 371 словоупотребление.

Таблица 1

Примеры ключевых слов, отобранных из текстов журнала «Future Generation Computer Systems» за 2016–2019 гг. (фрагмент таблицы)

Слово	Показатели частоты (F) и частотности (I_f, %) в подкорпусе								Средний показатель частотности, %
	2016		2017		2018		2019		
	F	I_f	F	I_f	F	I_f	F	I_f	
algorithm	175	0,3	280	0,4	808	0,5	718	0,3	0,4
application	249	0,4	317	0,4	609	0,3	740	0,3	0,4
cloud	508	0,9	511	0,7	1083	0,6	853	0,4	0,6
compute	250	0,4	333	0,5	766	0,4	683	0,3	0,4
datum	584	1	814	1,2	1671	1	1821	0,9	1

Представленная табл. 1 была получена путем объединения ключевых слов, отобранных в результате статистического анализа каждого подкорпуса текстов. Далее проводился анализ параметра частотности для каждого полученного ключевого слова во всех подкорпусах текстов с целью определения изменения его частотности в диахроническом аспекте.

Согласно полученным данным, 60 % слов (например, *algorithm, application, cloud, compute, datum, model, performance, service, system*) являются ключевыми во всех 4 исследуемых подкорпусах текстов. 20 % слов (*approach, Internet, network, resource*) имеют показатель частотности $\geq 0{,}3$ % в 3 подкорпусах текстов, 5 % слов (*process*) являются ключевыми в 2 подкорпусах текстов и 15 % слов (*schedule, smart, task*) были определены как ключевые только в одном подкорпусе текстов.

В целом 15 % лексических единиц (*internet, network, propose*) характеризуются увеличением частотности употребления, а 30 % единиц (*application, approach, cloud, resource, schedule, service*) – снижением данного показателя. Относительной устойчивостью обладает 25 % лексических единиц (*model, paper, performance, system, base)*; волнообразное изменение частотности в диахроническом плане свойственно 30 % слов (*algorithm, compute, datum, process, smart, task)*.

Анализ среднего показателя параметра частотности позволил выделить 17 ключевых слов из 20 отобранных.

На следующем этапе был проведен статистический анализ 4 подкорпусов текстов журнала «Information and Software Technology» (табл. 2): 2016 г. – 33 945 словоупотреблений, 2017 г. – 29 184 словоупотребления, 2018 г. – 43 854 словоупотребления, 2019 г. – 41 014 словоупотреблений.

Таблица 2

Примеры ключевых слов, отобранных из текстов журнала «Information and Software Technology» за 2016–2019 гг. (фрагмент таблицы)

Слово	Показатели частоты (F) и частотности (I_f, %) в подкорпусе								Средний показатель частотности, %
	2016		2017		2018		2019		
	F	I_f	F	I_f	F	I_f	F	I_f	
base	105	0,3	101	0,3	97	0,2	104	0,2	0,3
code	54	0,1	106	0,3	151	0,3	160	0,3	0,3
software	436	1,2	346	1,1	647	1,4	553	1,3	1,3
system	158	0,4	180	0,6	199	0,4	186	0,4	0,5
test	151	0,4	118	0,4	205	0,4	181	0,4	0,4

Среди представленных программой ключевых слов 38 % лексических единиц (*application, approach, method, model, process, propose, software, study, system, test)* имеют показатель частотности ≥ 0,3 % во всех 4 подкорпусах текстов. В 3 подкорпусах ключевыми являются 12 % слов (*code, development, project*); 31 % слов (*analysis, base, engineering, paper, requirement, research, technique, testing*) дважды зафиксированы как частотные; 19 % единиц являются ключевыми только в выборке за один год. Например, *datum* и *mutation* оказываются ключевыми только в текстах 2017 г., *performance* – 2018 г., а *identify* и *quality* – 2019 г.

В табл. 2 также отмечается ряд тенденций, касающихся изменения частотности отобранных лексических единиц. Было выявлено 58 % слов, которые относительно равномерно представлены в текстах журналов за все годы (например, *analysis, application, method, model, project, quality, research, test*). Волнообразная частотность зафиксирована в 31 % случаев, например, *datum, development, performance, requirement, software, system*. Кроме того, выявлено два случая повышения частотности (*code* и *technique*) и один случай снижения частотности (*process*).

В результате подведения итогов по всем четырем подкорпусам текстов рассматриваемого журнала было установлено, что 77 % слов имеют показатель частотности 0,3 % и более. Следовательно, в итоговый список ключевых слов было внесено 20 лексических единиц из 26 отобранных.

На третьем этапе были проанализированы тексты журнала «Journal of Information Technology» (табл. 3): 2016 г. – 4 106 словоупотреблений, 2017 г. – 4 731 словоупотребление, 2018 г. – 4 087 словоупотреблений, 2019 г. – 3 345 словоупотреблений.

Таблица 3

Примеры ключевых слов, отобранных из текстов журнала «Journal of Information Technology» за 2016–2019 гг. (фрагмент таблицы)

Слово	Показатели частоты (F) и частотности (I_f, %) в подкорпусе								Средний показатель частотности, %
	2016		2017		2018		2019		
	F	I_f	F	I_f	F	I_f	F	I_f	
digital	7	0,1	5	0,1	35	0,8	12	0,3	0,3
information	15	0,3	33	0,6	65	1,5	43	1	0,9
model	17	0,4	10	0,2	22	0,5	8	0,1	0,3
system	20	0,4	15	0,3	15	0,3	49	1,1	0,5
technology	15	0,3	37	0,7	16	0,3	13	0,3	0,4

Результаты анализа полученных данных показали, что 67 % отобранных слов являются ключевыми в одном подкорпусе текстов, например, *architecture, control, design, electronic, sourcing*. В полученной выборке отсутствуют слова, являющиеся ключевыми для 3 подкорпусов текстов. При этом только 11 % слов (*information, research, social, study, system, technology*) обладают высокой частотностью употребления в текстах за все четыре года; 23 % единиц имеют показатель, равный или превышающий 0,3 % в публикациях за два года, например, *digital, knowledge, model, platform.*

Следует отметить, что лексическое наполнение текстов рассматриваемого журнала характеризуется вариативностью в зависимости от года издания. В связи с этим в первоначальный список ключевых слов были включены 57 лексических единиц, при этом 9 % из них являются ключевыми для одного подкорпуса текстов, а в других подкорпусах отсутствуют (2016 г. – *resistance*, 2017 г. – *crowdfunding*, 2018 г. – *stock, TML*, 2019 г. – *architecture, overload)*.

Диахронический анализ частотности слов, которые встречались как минимум в двух подкорпусах текстов журнала показал, что только 2 % слов (*process*) равномерно представлены с 2016 г. по 2019 г. В 14 % случаев отмечается повышение частотности употребления (*analytics, architecture, dynamic, electronic, overload, relationship*), а в 7 % случаев было выявлено снижение частотности употребления слов (*factor, paper, resistance, review*). Наиболее многочисленную группу слов (79 %) составляют те лексические единицы, которым свойственно волнообразное изменение частотности употребления с течением времени, например, *control, datum, digital, system*. По результатам анализа параметра частотности было выделено 16 ключевых слов из 57 отобранных раннее.

Как показывает анализ 12 подкорпусов текстов журналов, в список ключевых слов вошли только те единицы, которые встретились в выборках текстов как минимум за два года. Следовательно, для достижения установ-

ленного показателя 0,3 % лексическая единица должна обладать высокой частотностью и определенной устойчивостью. Вхождение лексической единицы в списки ключевых слов за весь исследуемый период может свидетельствовать о том, что она является одним из базовых концептов предметной области ИТ или клише научного стиля. Обнаружение слова только в одном подкорпусе текстов свидетельствует о неустойчивом интересе исследователей к понятию, обозначаемому данным словом. Можно предположить, что фиксация снижения, увеличения или волнообразного изменения частотности слов связана с изменением потребностей в изучении каких-либо технологий в силу их досконального или неполноценного исследования или недостаточного обеспечения теми или иными средствами, необходимыми для исследования.

Отдельного внимания заслуживают лексические единицы, которые не вошли в список ключевых слов, но характеризуются увеличением показателя частотности в выборке за 2019 г.: *analytics, architecture, dynamic, electronic, identity, overload, quality, relationship, technique*. Такая тенденция может свидетельствовать о возможном сохранении и увеличении частотности их употребления в ходе развития предметной области ИТ, в связи с чем представляется целесообразным на данном этапе внести рассмотренные 9 лексических единиц в список ключевых слов и в дальнейшем провести их семантический анализ.

На четвертом этапе был проведен статистический анализ 8 подкорпусов текстов руководств по эксплуатации программных продуктов Microsoft (табл. 4). Представленные в табл. 4 подкорпусы содержат следующее количество словоупотреблений: Excel – 37 024; OneDrive – 5 324; SharePoint – 2 734; Word – 11 245.

Таблица 4

Примеры ключевых слов, отобранных из текстов руководств компании Microsoft (фрагмент таблицы)

Слово	Показатели частоты (F) и частотности (I_f, %) в подкорпусе							
	Excel		OneDrive		SharePoint		Word	
	F	I_f	F	I_f	F	I_f	F	I_f
content	24	0	–	–	11	0,4	51	0,4
field	127	0,3	–	–	–	–	37	0,3
option	113	0,3	–	–	10	0,3	46	0,4
pane	21	0	5	0	9	0,3	23	0,2
sync	3	0	24	0,4	11	0,4	–	–

Следует отметить, что в первоначальной выборке высокой частотностью обладали лексические единицы, обозначающие названия рассматриваемого

программного обеспечения, например *Excel, Microsoft Team, Office*. Данные единицы были исключены из списка полученных ключевых слов, поскольку в рамках проводимого исследования было принято решение не рассматривать номены. По результатам статистической обработки рассматриваемых текстов было получено 119 ключевых слов. Несмотря на разноплановость информации, предоставляемой в руководствах по эксплуатации разных программ, было выявлено 39 % слов, являющихся ключевыми для нескольких подкорпусов текстов. При этом 61 % отобранных слов являются ключевыми для одного подкорпуса текстов.

На заключительном этапе ключевые слова, отобранные из 4 источников (3 периодических изданий и руководства по эксплуатации), были сведены в общий перечень ключевых слов исследуемых современных текстов по ИТ, который включает 166 лексических единиц, при этом 2,4 % ключевых слов содержатся в трех источниках из четырех отобранных, 4,2 % – в двух источниках. Значительное количество ключевых слов (93,4 %) зафиксированы только в одном источнике, при этом не было выявлено ни одного слова, которое является ключевым во всех источниках. Такие статистические данные позволяют сделать вывод о том, что информация, представленная в журналах и руководствах, разнообразна по своему содержанию. Следовательно, подтверждается обширность охватываемых в рамках ИТ тем и вопросов и обосновывается необходимость включения в корпус текстов для исследования источников разного характера. Кроме того, проверка отобранных ключевых слов по толковым словарям [6; 7; 8; 9; 10; 11; 12] показала, что 94 % лексических единиц имеют зафиксированные дефиниции, что подтверждает высокую эффективность применения автоматизированных средств для оперативного извлечения ключевых слов.

ЛИТЕРАТУРА

1. *Зубов, А. В.* Способы автоматического извлечения терминов из текстов / А. В. Зубов // Slavonic Terminology Today. – Vol. CLXVII, Book 28. – Belgrade, 2017. – С. 639–642.
2. Future Generation Computer Systems [Electronic resource] // Elsevier. – Mode of access: https://www.journals.elsevier.com/future-generation-computer-systems. – Date of access : 5.12.2019.
3. Information and Software Technology, Supports open access [Electronic resource] // Elsevier. – Mode of access : https://www.sciencedirect.com/journal/information-and-software-technology/vol/116/suppl/C. – Date of access : 5.12.2019.
4. Journal of Information Technology [Electronic resource] // Palgrave Macmillan. – Mode of access : https://www.palgrave.com/gp/journal/41265/volumes-issues/latest-issue. – Date of access : 5.12.2019.
5. Office 365 Training Center [Electronic resource] // Microsoft. – Mode of access : https://support.office.com/en-gb/office-training-center. – Date of access : 28.11.2019.

6. *Butterfield, A.* Oxford Dictionary of Computer Science / A. Butterfield, G. E. Ngondi, A. Kerr. – Oxford : Oxford Univ. Press, 2016 – 627 p.
7. *Collin, S. M. H.* Dictionary of Computing / S. M. H. Collin. – 5th ed. – London : Bloomsbury, 2004. – 345 p.
8. *Collin, S. M. H.* Dictionary of ICT / S. M. H. Collin. – 4th ed. – London : Bloomsbury, 2004. – 288 p.
9. Dictionary of Computer and Internet Terms / D. A. Downing [et al.]. – 10th ed. – N. Y. : Barron's, 2009. – 560 p.
10. Information technology. Vocabulary : ISO/IEC 2382:2015(en) [Electronic resource] // Online Browsing Platform (OBP). – Mode of access : https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en. – Date of access : 4.09.2020.
11. Microsoft Computer Dictionary – 5th ed. – Redmond : Microsoft Press, 2002. – 637 p.
12. Dictionary of Computer and Internet Terms: in 2 vol. / ed. J. C. Rigdon. – Cartersville : Eastern Digital Resources, 2016. – 1 vol.

This article is devoted to the problem of automatic key word extraction from scientific texts based on the frequency criterion. As a result, 166 key words are obtained. The analysis showed that 93,4 % of the keywords are recorded in only one group of texts, which confirms the need for processing various types of sources and using automated tools for prompt extraction of keywords.

Поступила в редакцию 27.10.20