Д. В. Степанова

АНАЛИЗ МЕТОДОВ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ТЕРМИНОВ ИЗ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

В условиях активной информатизации всех сфер общественной жизни и необходимостью оперативной обработки значительных объемов научно-технической информации, поступающей на иностранных языках, исследования по автоматическому выделению терминов в настоящее время представляются особенно актуальными. Выделение терминов и терминологических словосочетаний из научно-технических текстов при решении задач обработки естественного языка является объектом изучения многих направлений прикладной лингвистики, например, в лексикографии при создании терминологических словарей, автоматического реферирования, аннотирования и индексирования текстов, рубрикации текстов и их тематической структуризации, перевода текстов, а также извлечения знаний из текстовых источников.

Практика перевода и составления терминологических словарей, а также последние теоретические исследования показывают, что одной из наиболее важных и сложных проблем, как при решении традиционных лингвистических задач, так и при создании информационно-поисковых словарей и тезаурусов является отбор терминов.

Вопрос об отборе терминов тесным образом связан с актуальными и дискуссионными проблемами терминоведения, а именно с проблемой определения термина и четких границ, отличающих термин от нетермина, а также методов выделения терминов из научно-технических текстов.

В нашем понимании термин представляет собой элемент общелитературного языка, специфика которого заключается в функции выражения научного понятия, занимающего определенное место в системе понятий некоторой области знания и в сфере его специального распространения [1; 2]. Функциональный подход к различению термина от нетермина заключается в соотнесенности термина с профессиональным, а нетермина с бытовым понятием. Следует отметить, что к терминам относятся не только слова, но и словосочетания, обозначающие единое, но расчлененное научное или техническое понятие [3; 4].

В специальной литературе предлагаются различные критерии определения терминологичности лексических единиц, например, дефинитивный критерий, критерий концептуальной целостности, семантический критерий, статистический критерий, критерий воспроизводимости, психолингвистический критерий.

На обязательность дефинированности терминов указывали В. П. Даниленко [5, с. 15], Т. Л. Канделаки [6, с. 3–11], Н. З. Котелова [7, с. 125], Э. Ф. Скороходько [8, с. 15] и другие исследователи. Под дефиницией понимается логическое определение понятия, при котором устанавливается содержание понятия (признаки предметов, отражаемых в понятии), его отличительные признаки [3, с. 62]. Дефиниция должна быть емкой, то есть содержать все необходимые и достаточные признаки обозначаемого понятия и отражать признаки, по которому можно отличить одно понятие от другого, потому что «внутренняя форма слова-термина является сложной и много-гранной категорией» [9, с. 17].

Согласно критерию концептуальной целостности, предложенному В. М. Овчаренко, в качестве термина должны признаваться синтетические или семантически целостные образования, выражающие специальные понятия, а не свободные, или аналитические сочетания знаков, которые объединяются по действующим языковым моделям для выражения таких понятий [10, с. 95].

В основе семантического критерия лежит методика логико-семантического соотнесения профессионального понятия и соответствующего ему языкового выражения с опорой на некоторые методические способы определения терминированности лексических единиц [11; 12].

Статистический критерий заключается в формальном выделении терминов без обращения к смыслу и с применением статистических методов. Данный способ может быть осуществлен при наличии статистических описаний по текстам исследуемого подъязыка и любого подъязыка, с которым производится сравнение. Эти описания представлены в виде алфавитно-частотных или частотно-алфавитных словарей [13, с. 7]. Мерой терминологичности языковой единицы в одном подъязыке относительно другого будет выступать величина разности номеров, или рангов, этой единицы в двух подъязыках. Незначительная разность номеров и сходные статистические характеристики в художественных и научно-технических текстах позволяют утверждать, что эти единицы являются общеупотребительными. Если разность для словоформ в научно-технических текстах невелика, но она заметна в художественных текстах, то эти слова или словоформы называются общенаучными терминами. Те лексические единицы, которые характеризуются большой разностью по отношению как к художественному, так и к научно-техническому тексту, называются специальными терминами [12, с. 250–255]. На основе данного подхода была также разработана система определения маркеров специальных областей деятельности [14].

Критерий воспроизводимости основывается на положении, что терминологические словосочетания в речи не производятся, а воспроизводятся [15].

В основе психолингвистического критерия лежит логико-интуитивный подход, который позволяет четко членить текст на непредикативные сегменты, однако затрудняет членение непредикативных терминологических сегментов, выражающих составное специальное понятие [16].

Н. Ю. Зайцева выделяет три группы критериев терминологичности, которые объединяют как признаки терминов, так и желательные или обязательные требования [17, с. 6–7]:

1) синтаксические (с точки зрения организации означающего терминологического знака);

2) семантические (с точки зрения организации означаемого терминологического знака);

3) прагматические или функциональные (с точки зрения употребления терминологического знака в соответствии с задачами коммуникации.

К синтаксическим критериям относятся: краткость, которая подразделяется на лексическую краткость или нетавтологичность – фиксация в форме термина минимального количества отличительных признаков, отсутствие элементов, не несущих смысловой нагрузки, и формальную краткость,

которая часто достигается путем создания эллиптических конструкций и аббревиатур; способность к деривации как важнейшему средству развития и обогащения терминологических систем; соответствие нормам языка, которое достигается устранением профессиональных жаргонизмов, отклонений от фонетических и грамматических норм, а также замещением несвойственных литературному языку форм на нормативные формы [17, с. 58–66].

Семантическая группа включает такие критерии, как обозначение понятия специальной области знания, непротиворечивость отражения терминов концептов понятия, системность (принадлежность обозначаемого термином понятия определенному месту в системе понятий данной предметной области), дефинированность, однозначность, точность значения, независимость смысла термина от контекста, стилистическая нейтральность, конвенциональность и целенаправленный характер появления термина, неизменчивость значения устоявшихся терминов, полнозначность, отсутствие синонимов, номинативность [17, с. 7–58].

К прагматическим критериям относятся: внедренность (общепринятость или употребительность термина или ТС), современность (соответствие терминологического знака уровню развития области знания), благозвучность и эзотеричность (намеренная недоступность, обусловленная желанием оградить профессиональное общение от его восприятия неспециалистами) [17, с. 67–70].

По мнению Н. Ю. Зайцевой наиболее важными являются семантические критерии, поскольку специфика терминов лежит не столько в плане выражения, сколько в плане содержания, в характере его значения. Исследователь выделяет три базовых критерия терминологичности: обозначение понятия специальной области знания, системность и дефинированность.

В целях решения задачи автоматического выделения новых терминов и терминологических словосочетаний из научно-технических текстов, а также для формализации такой процедуры для существующих терминов, специалистами разработано и предлагается достаточное количество методов.

Существующие методы классифицируются по следующим критериям: адаптивность, количество слов в извлекаемых терминах, используемые лингвистические ресурсы, применяемые количественные методы, способ фильтрования слов общей лексики [18].

Адаптивные методы автоматического выделения терминов основываются на текстах, которые, с одной стороны, служат для определения правил извлечения терминов, а с другой, являются базой для выделения терминов, что гарантирует достаточно высокие результаты только при работе с данными текстами.

Как известно, статистические методы достаточно эффективны при выделении однословных терминов и терминологических словосочетаний, состоящих их двух компонентов. Определенную сложность представляет собой выделение терминов произвольной длины. Это обусловлено проблемой определения границ терминов. В некоторых случаях для решения данной проблемы, в частности, для выделения именных терминологических словосочетаний может быть использован синтаксический анализатор, главная функция которого заключается в как можно быстрой и точной идентификации основных составляющих (и возможно синтаксических функций) в исходном тексте [19; 20].

Решение этой задачи может осуществляться несколькими путями.

1. Использование маркеров границ термина. На этом этапе текст разбивается на именные словосочетания максимальной длины при помощи лексико-семантических средств, определяющих границы этого типа словосочетаний (глаголы, местоимения, предлоги, определители).

2. Определение следования слов определенных частей речи, в которых может употребляться термин. Конструкции, которые совпадают с заранее заданными словосочетаниями, признаются терминологическими словосочетаниями.

3. Использование синтаксического анализатора для получения информации, необходимой для выделения словосочетаний. По результатам экспериментов данные, полученные анализатором, не являются высокоточными.

Кроме того, для выделения терминологических словосочетаний предлагается процедура определения цепочки слов на основании списка слов и знаков, которые не могут входить в состав терминологического словосочетания [21]. Далее происходит вычисление веса термина-кандидата в зависимости от частоты его самостоятельного употребления и употребления в составе словосочетаний. При этом, чем выше частота употребления в составе других словосочетаний, тем ниже его вес. Такой метод позволяет выделить многокомпонентные терминологические словосочетания.

Автоматическое выделение терминов может также быть основано на данных лингвистического анализа [18]. Выделяют четыре основных уровня данного анализа:

1) графематический анализ, который, например, позволяет выделить адреса электронной почты, даты, ссылки на электронные документы;

2) морфологический анализ, данные которого могут послужить основой для выделения шаблонов терминов, например, именных или глагольных групп;

3) синтаксический анализ, который является основой для поиска лексико-синтаксических шаблонов и морфосинтаксических связей;

4) семантический анализ, который предполагает использование справочной литературы в качестве словарной базы анализа.

Наряду с рассмотренным выше статистическим критерием определения терминологичности лексических единиц, специалисты также предлагают метод рекурсивного вхождения терминов в более длинные термины [21].

Для автоматического выделения терминов и терминологических словосочетаний применяются также статистические и семантические фильтры [22; 23; 24].

Интерес к разработке и использованию статистических фильтров был вызван тем фактом, что частота употребления терминов в специальном контексте отличается от частоты употребления нетерминов. Статистические характеристики могут варьироваться от самых простых (показатель частотности) до сложных (получение данных при помощи формул). Следует отметить, что в условии небольшого объема корпуса и в результате невысокой частотности употребления единиц, данные сложных фильтров являются приблизительными. Результаты статистических фильтров зависит от того, насколько точно подобран определенный фильтр к данному тексту.

Идея семантических фильтров была предложена исследователями недавно, но уже получила широкое распространение. В основе разработки таких фильтров лежит положение о том, что термины выражают и формируют специальные понятия и, таким образом, имеют определенные семантические характеристики, отличающие их от нетерминов. Эксперимент показал, что применение семантического фильтра увеличила точность полученных данных на 5–9 %.

Анализ существующих методов автоматического выделения терминов из научно-технических текстов показывает, что, несмотря на все положительные стороны и возможности рассмотренных методов, некоторые проблемы остаются нерешенными, и, как следствие, пользователи получают не всегда точный и полный результат. При этом, одним из способов проверки работы того или иного метода является сопоставление полученных результатов и мнения эксперта в исследуемой специальной области знания. В результате, в настоящее время для успешного выделения терминов из научно-технических текстов наряду с использованием современных методов следует учитывать и мнение лингвиста и специалиста определенной сферы деятельности.

ЛИТЕРАТУРА

1. *Винокур, Г. О.* О некоторых явлениях словообразования в русской технической терминологии / Г. О. Винокур // «Труды МИФЛИ» : сб. ст. по языковедению. – М., 1939. – С. 3–54.
2. *Головин, Б. Н.* О некоторых проблемах изучения терминов / Б. Н. Головин // Научный симпозиум «Семиотические проблемы языков науки, терминологии и информатики» : матер. симпозиума; Москва, 1971 г.: в 2 ч.; редкол. : А. Г. Волков (отв. за вып.) [и др.]. – М. : Изд-во Моск. ун-та, 1971. – Ч. 1. – С. 64–67.
3. *Головин, Б. Н.* Лингвистические основы учения о терминах / Б. Н. Головин, Р. Ю. Кобрин. – М. : Высш. шк., 1987. – 104 с.
4. *Хроменков, П.* Современные системы машинного перевода: общие черты и различия / П. Хроменков // Компьютерная лингвистика и обучение языкам : сб. науч. ст. / МГЛУ; редкол.: А. В. Зубов (отв. ред.) [и др.]. – Минск : МГЛУ, 2000. – С. 154–159.
5. *Даниленко, В. П.* Русская терминология: Опыт лингвистического описания / В. П. Даниленко; АН СССР, Ин-т рус. яз. – М. : Наука, 1977. – 246 с.
6. *Канделаки, Т. Л.* Значение терминов и системы значений научно-технических терминологий / Т. Л. Канделаки // Проблемы языка науки и техники. Логические, лингвистические и историко-научные аспекты терминологии : сб. ст.; отв. ред. чл.-кор. АН СССР С.Г. Бархударов. – М. : Наука, 1970. – С. 3–39.
7. *Котелова, Н. З.* К вопросу о специфике термина / Н. З. Котелова // Лингвистические проблемы научно-технической терминологии: матер. совещания, АН СССР г. Ленинград 30 мая–2 июня 1967 г.; редкол.: С. Г. Бархударов (отв. ред.) [и др.]. – М. : Наука, 1970. – С. 122–126.
8. *Скороходько, Э. Ф.* Вариативность и скрытая полисемия в терминологии / Э. Ф. Скороходько // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – М., 2003. – № 4. – С. 15–19.
9. *Кияк, Т. Р.* Лингвистические аспекты терминоведения / Т. Р. Кияк. – Киев : УМКВО, 1989. – 103 с.

10. *Овчаренко, В. М.* Термин, аналитическое наименование и номинативное определение / В. М. Овчаренко // Современные проблемы терминологии в науке и технике : сб. ст.; отв. ред. В. С. Кулебакин. – М. : Наука, 1969. – С. 91–121.
11. *Головин, Б. Н.* О некоторых доказательствах терминированности словосочетаний / Б. Н. Головин // Лексика, терминология, стили : межвуз. науч. сб.; Горьков. гос. ун-т им. Н. И. Лобачевского; редкол. : Б. Н. Головин (отв. ред.) [и др.]. – Горький : ГГУ, 1973. – Вып. 2. – С. 57–65.
12. *Пиотровский, Р. Г.* Статистическое опознание термина / Р. Г. Пиотровский, С. В. Ястребова // Статистика текста : сб. ст.; редкол.: А. И. Киселевский (гл. ред.) [и др.]. – Минск : БГУ, 1969. – С. 249–259.
13. *Зубов, А. В.* Основы лингвистической информатики : учеб. пособие : в 3 ч. / А. В. Зубов, И. И. Зубова / Минск. гос. пед. ин-т ин. яз.; редкол. : А. Д. Борисевич, С. А. Истомин. – Минск. : МГПИИЯ, 1992. – Ч. 2. – 138 с.
14. *Le An Ha.* A Practical Comparison of Different Filters Used in Automatic Term Extraction / An Ha Le // Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, 26–28 May 2004. – Lisbon, 2004. – P. 511–514.
15. *Кобрин, Р. Ю.* О сопоставительном анализе свободных фразеологических и терминологических сочетаний / Р. Ю. Кобрин // Ученые записки Горьковского гос. ун-та. Серия лингвистическая : матер. и исследов. по рус. грамматике и лексикологии / Горьк. гос. ун-т им. Н.И. Лобачевского; редкол.: Н. Д. Русинов (отв. ред.) [и др.]. – Горький : ГГУ, 1970. – Вып. 99. – С. 15–28.
16. *Кобрин, Р. Ю.* Психолингвистический эксперимент для оценки терминологичности элементов текста / Р. Ю. Кобрин // Автоматическая переработка текста методами прикладной лингвистики. – Кишинев, 1977. – С. 11–12.
17. *Зайцева, Н. Ю.* Информационно-семиотическая природа термина и типология языков / Н. Ю. Зайцева. – СПб. : Изд-во РГПУ им. А. И. Герцена, 2002. – 84 с.
18. *Табарча, А. И.* Анализ и сравнение методов автоматического извлечения терминов из текста / А. И. Табарча // Аспирант и соискатель. – 2010. – № 6. – С. 133–137.
19. *Hulth, A.* Improved Automatic Keyword Extraction Given More Linguistic Knowledge / A. Hulth // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), Sapporo, 11–12 July, 2003. – Sapporo, 2003. – P. 216–223.
20. *Kinyon, A.* A language-Independent Shallow Parser Compiler / A. Kinyon // Proceedings of the Annual Meeting of the Association for Computational Linguistics, Toulouse, 6–11 July, 2001. – Toulouse, 2001. – P. 322–329.
21. *Браславский, П.* Сравнение пяти методов извлечения терминов произвольной длины / П. Браславский, Е. Соколов // Матер. Межд. конф. Диалог'2008 «Компьютерная лингвистика и интеллектуальные технологии». – М. : РГГУ, 2008. – С. 67–74.
22. *Le An Ha.* A Practical Comparison of Different Filters Used in Automatic Term Extraction / An Ha Le // Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, 26–28 May 2004. – Lisbon, 2004. – P. 511–514.
23. *Maynard, D.* Identifying Contextual Information for Multi-Word Term Extraction / D. Maynard, S. Ananiadou // Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering TKE-99, Innsbruck, 25–27 August, 1999. – Vienna : TermNet – Verlag, 1999. – P. 212–221.
24. *Paice, C. D.* A Three-Prolonged Approach to the Extraction of Key Terms and Semantic Roles / C. D. Paice, W. J. Black // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), Sapporo, 11–12 July, 2003. – Sapporo, 2003. – P. 357–363.