DOI: 10.37791/2687-0649-2021-16-6-21-27

Метод извлечения русскоязычных многокомпонентных терминов из научно-технических текстов

Ю. И. Бутенко[1*]***, Ю. В. Строганов***[1]***, А. М. Сапожков***[1]
[1] *Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия*
** iubutenko@bmstu.ru*

Аннотация. В статье представлен метод извлечения русскоязычных многокомпонентных терминов из научно-технических текстов на основе структурных моделей терминологических словосочетаний. Описаны существующие подходы к извлечению терминов на основе метода извлечения устойчивых словосочетаний, статистических и гибридных методов, а также отмечены лингвистические аспекты терминоведения, не охваченные перечисленными методами. Охарактеризован лексический состав научно-технических текстов, приведена классификация специальной лексики в научно-технических текстах. Изучены структурные особенности терминологической лексики. Представлены наиболее продуктивные модели многокомпонентных терминологических словосочетаний в русском языке. Предложен метод извлечения русскоязычных многокомпонентных терминов из научно-технических текстов, а также описаны его этапы. Показано, что на первом этапе проводится морфолого-синтаксический анализ текста путем приписывания каждому слову его грамматических характеристик. Затем происходит исключение частей речи, которые не могут входить в состав русскоязычных многокомпонентных терминов, а также стоп-слов, которые вместе с термином образуют свободные словосочетания. Полученные цепочки слов далее соотносятся с шаблонами терминологических словосочетаний, имеющихся в базе структурных моделей терминов, а также с терминологическим словарем на предмет наличия исследуемого термина-кандидата. Обоснована необходимость привлечения терминолога для разрешения неоднозначных случаев. Каждый этап метода извлечения русскоязычных многокомпонентных терминов из научно-технических текстов проиллюстрирован примерами. Перечислены перспективы исследования, а также обоснована необходимость усложнения методов извлечения терминов путем дальнейшей классификации терминологической лексики по формальной и семантической структурам, видам антропоморфных терминов, номенклатурным названиям, нормативности/ненормативности терминологических единиц.

Ключевые слова: корпус текстов, научно-технические тексты, извлечение терминов, структура научно-технического текста, многокомпонентный термин

Для цитирования: *Бутенко Ю. И., Строганов Ю. В., Сапожков А. М.* Метод извлечения русскоязычных многокомпонентных терминов из научно-технических текстов // Прикладная информатика. 2021. Т. 16. № 6. С. 21–27. DOI: 10.37791/2687-0649-2021-16-6-21-27

Method for the extraction of Russian-language multicomponent terms from scientific and technical texts

Iu. Butenko[1*]***, Yu. Stroganov***[1]***, A. Sapozhkov***[1]
[1] *Bauman Moscow State Technical University, Moscow, Russia*
** iubutenko@bmstu.ru*

Abstract. The article presents a method for extracting Russian-language multicomponent terms from scientific and technical texts based on structural models of terminological collocations. The existing approaches to term extraction on the basis of the method of stable word combination extraction, statistical and hybrid methods are described, and the linguistic aspects of terminology, not covered by the listed methods, are noted. The lexical composition of scientific and technical texts is characterized, the classification of special vocabulary in scientific and technical texts is given. The structural features of terminological vocabulary have been studied. The most productive models of multi-component terminological word combinations in Russian are presented. A method for extracting Russian-language multicomponent terms from scientific and technical texts is offered, and its stages are described. It is shown that the first stage involves morphological and syntactic analysis of the text by attributing to each word its grammatical characteristics. Then there is the exclusion of parts of speech, which can not be part of the Russian multisyllabic terms, as well as stop-words, which together with the term form free word combinations. The resulting word chains are further correlated with the templates of terminological word combinations available in the database of structural models of terms, as well as the terminological dictionary for the presence of the studied candidate term. The necessity of involving a terminologist to resolve ambiguous cases is substantiated. Each step of the method for extracting Russian-language multicomponent terms in scientific and technical texts is illustrated by examples. Further research perspectives are listed, and the necessity of complicating the methods of text extraction, by further classification of terminological vocabulary according to formal and semantic structures, types of anthropomorphic terms, nomenclatural names, normativity/non-normativity of terminological units is substantiated.

Keywords: text corpus, scientific and technical texts, term extraction, structure of scientific and technical text, multi-component term

For citation: Butenko Iu., Stroganov Yu., Sapozhkov A. Method for the extraction of Russian-language multicomponent terms from scientific and technical texts. *Prikladnaya informatika*=Journal of Applied Informatics, 2021, vol.16, no.6, pp.21-27 (in Russian). DOI: 10.37791/2687-0649-2021-16-6-21-27

Введение

Корпуса текстов обычно размечаются для удобства пользования, т. е. текстам и содержащимся в них языковым единицам приписываются специальные метки. Размеченные корпуса обеспечивают специализированными поисковыми системами, реализующими грамматические и лексические виды поиска [4]. Так, для корпуса научно-технических текстов наибольшую значимость приобретает терминологическая разметка, так как именно термины выступают основным средством передачи информации [9].

Работа с корпусами научно-технических текстов требует особого инструментария для выявления устойчивых термино-

логических сочетаний [13]. Среди наиболее распространенных методов выявления многокомпонентных терминов в текстах используют метод выявления устойчивых сочетаний, статистический и гибридный методы [1]. В основе метода выявления устойчивых сочетаний лежит использование грамматики лексико-синтаксических шаблонов, представляющих собой структурные модели лингвистических конструкций [12]. Статистический подход заключается в нахождении n-грамм по заданным частотным характеристикам. Гибридный подход для выделения терминологических сочетаний, объединяющий лингвистический и статистический методы, заключается в предварительном описании моделей, по которым могут быть построены термины, для последующего нахождения их в корпусе [2, 3]. Внутри множеств однотипных синтаксических конструкций выполняется ранжирование в соответствии с той или иной статистической мерой [11]. В рамках указанных методов отсутствует возможность отличить многокомпонентный термин от свободного словосочетания, например *неорганическая химия* и *современная химия* [10]. В текстах часто встречаются и словосочетания вида «общеупотребительное слово и многокомпонентный термин», например *современная неорганическая химия*. В то же время развивающиеся терминологии компьютерных наук, авиации и космонавтики, нанотехнологий и других предметных областей используют новые способы терминообразования, таким образом расширяя структурные модели терминов [8].

Целью исследования является описание метода извлечения русскоязычных многокомпонентных терминологических единиц из корпуса научно-технических текстов на основе структурных моделей терминологических словосочетаний, дополненных грамматическими и лексическими ограничениями на сочетаемость слов в составе многокомпонентного термина.

Формальная структура многокомпонентных терминов

В научно-технических текстах используется специальная лексика следующих видов: общенаучная терминология, узкоспециализированные термины и номенклатуры. Общенаучная терминология представлена собирательными наименованиями, которые могут быть названиями терминологий, например *анализ, синтез, вид, метод, процесс, материал* и др., и отражает абстрактные понятия, зачастую входящие в состав дефиниций. Но с другой стороны, общенаучная лексика входит в состав многокомпонентных терминологических словосочетаний, которые обозначают одно понятие, например *методы химического анализа, оборудование для химического анализа, виды сварки, сварочный материал*. Номенклатурой называют терминологическое обозначение частного специального понятия какой-либо предметной области, например *Ту-156, Уолл Стрит, Захват-22*.

Термин может быть однокомпонентным и состоять из ключевого слова или представлять собой терминологическую группу, в состав которой входит ключевое слово или ядро группы, одно или несколько левых определений и одно или несколько правых или предложных определений, которые уточняют или модифицируют смысл терминологической единицы [5].

Наиболее сложным явлением в процессе автоматического извлечения терминов из научно-технических текстов представляют собой многокомпонентные термины – терминологические словосочетания, образованные лексическим и синтаксическим способами, то есть словосочетания, образованные по определенным моделям [6]. В основе анализа терминологических словосочетаний лежит вычленение исходного терминологического словосочетания и определение последовательности присоединения к нему остальных элементов. Исходным терминологическим словосочетанием, как пра-

вило, является двухкомпонентное субстантивное терминологическое словосочетание, которое в рамках трех-, четырехкомпонентного терминологического словосочетания характеризуется более тесными структурно-семантическими отношениями [5].

Наиболее продуктивным способом номинации является образование составных терминов, состоящих из двух, трех, четырех и пяти компонентов [5]. На практике встречаются термины, состоящие из 10–11 компонентов. В таблице 1 представлены наиболее продуктивные модели многокомпонентных терминологических словосочетаний в русском языке.

Таблица 1. Основные структурные модели многокомпонентных терминологических словосочетаний

Table 1. Basic structural models of multi component term combinations

Количество компонент *Number of components*	**Структурные модели** *Structural models*
1	S_1
2	$(S_6 \wedge S_1) \vee (S_9 \wedge S_1) \vee (S_1 \wedge S_2) \vee (S_6 \wedge S_1)$
3	$(S_6 \wedge S_6 \wedge S_1) \vee (S_6 \wedge S_1 \wedge S_2) \vee (S_6 \wedge S_1 \wedge S_8) \vee (S_6 \wedge S_1 \wedge S_3)$
4	$(S_6 \wedge S_1 \wedge S_{11} \wedge S_5) \vee (S_6 \wedge S_1 \wedge S_{11} \wedge S_4)$

Множество S соответствует морфологическим характеристикам слов русского языка, которые входят в состав многокомпонентного словосочетания. Множество S состоит из 12 элементов: S_1 – имя существительное в именительном падеже, S_2 – имя существительное в родительном падеже, S_3 – имя существительное в винительном падеже, S_4 – имя существительное в творительном падеже, S_5 – имя существительное в предложном падеже, S_6 – имя прилагательное, S_7 – имя числительное, S_8 – наречие, S_9 – причастие, S_{10} – деепричастие, S_{11} – предлог. В связи с тем, что морфологические анализаторы не всегда корректно и полно определяют грамматические характеристики имен прилагательных S_6, в предложенных моделях для них используется только характеристика части речи, а категории рода, числа и падежа будут такие же, как у имени существительного, которое стоит после имени прилагательного.

Добавление структурных моделей терминологических словосочетаний с пятью и более компонентами является нецелесообразным в силу низкой частотности употребления таких терминов на практике.

Метод извлечения многокомпонентных терминов из научно-технических текстов

Предлагаемый метод автоматического извлечения русскоязычных многокомпонентных терминов на основе базы данных структурных моделей терминологических словосочетаний состоит из пяти основных этапов. В качестве примера рассмотрим извлечение терминов для следующего фрагмента текста: *Как наука химия природных соединений возникла одновременно с органической химией.*

1. На первом этапе проводим анализ предложения по частям речи, например:

Как $_{\text{союз}}$ наука $_{\text{СУЩ,неод,жр ед,им}}$ химия $_{\text{СУЩ,неод,жр ед,им}}$ природных $_{\text{ПРИЛ}}$ соединений $_{\text{СУЩ,неод,ср мн,рд}}$ возникла $_{\text{ГЛ,сов,неперех жр,ед,прош,изъяв}}$ одновременно $_{\text{НАР}}$ с $_{\text{ПРЕДЛ}}$ органической $_{\text{ПРИЛ}}$ химией $_{\text{СУЩ,неод жр ед,тв}}$.

2. В состав терминологических словосочетаний не входят глаголы, союзы, местоимения, частицы, знаки препинания, а также некоторые сочетания частей речи, такие как «наречие + предлог» и прочее.

~~Как $_{\text{союз}}$~~ наука $_{\text{СУЩ,неод,жр ед,им}}$ химия $_{\text{СУЩ,неод,жр ед,им}}$ природных $_{\text{ПРИЛ}}$ соединений $_{\text{СУЩ,неод,ср мн,рд}}$ ~~возникла $_{\text{ГЛ,сов,неперех жр,ед,прош,изъяв}}$~~ ~~одновременно $_{\text{НАР}}$~~ ~~с $_{\text{ПРЕДЛ}}$~~ органической $_{\text{ПРИЛ}}$ химией $_{\text{СУЩ,неод,жр ед,тв}}$.

Таким образом, остаются следующие цепочки слов:

наука $_{\text{СУЩ,неод,жр ед,им}}$

химия $_{\text{СУЩ,неод,жр ед,им}}$ природных $_{\text{ПРИЛ}}$ соединений $_{\text{СУЩ,неод,ср мн,рд}}$

органической $_{\text{ПРИЛ}}$ химией $_{\text{СУЩ,неод,жр ед,тв}}$.

3. Проверяем полученные термины-кандидаты на наличие стоп-слов, указанных в специальной зоне словаря, и если они есть, убираем их. Под стоп-словами понимаем слова, которые образуют широко используемые коллокации с терминами, но в совокупности не являются терминами по сути, например *современная* химия, *рассматриваемый* метод синтеза о-гликозидов.

4. Полученные цепочки слов соотносим с шаблонами терминологических словосочетаний, имеющихся в базе структурных моделей терминов:

наука $_{\text{СУЩ,неод,жр ед,им}}$ – S_1 принадлежит к моделям терминов;

химия $_{\text{СУЩ,неод,жр ед,им}}$ природных $_{\text{ПРИЛ}}$ соединений $_{\text{СУЩ,неод,ср мн,рд}}$ – $S_1 \wedge S_6 \wedge S_2$ принадлежит к моделям терминов;

органической $_{\text{ПРИЛ}}$ химией $_{\text{СУЩ,неод,жр ед,тв}}$ – $S_6 \wedge S_1$ принадлежит к моделям терминов.

5. Полученные термины-кандидаты проверяем по словарю корпуса.

5.1. Если термин-кандидат есть в словаре, то извлекаем его как термин.

5.2. Если полученный термин-кандидат отсутствует в словаре, то отправляем его терминологу для обработки вручную.

5.3. Если термин-кандидат состоит из нескольких слов, то пробуем разбить его на несколько терминов.

Исследования терминологических словосочетаний в испанском, английском, немецком и других языках свидетельствуют о наличии структур многокомпонентных терминов, образованных по определенным моделям, отражающим особенности языка, на котором они образованы [7]. Для использования описанного метода извлечения многокомпонентных терминологических словосочетаний из научно-технических текстов на разных языках необходимо только наличие базы данных структурных моделей терминологических словосочетаний обрабатываемого языка.

Перспективой дальнейшего исследования также является усовершенствование метода извлечения многокомпонентных терминов путем приписывания дополнительных классификационных тегов к термину:

- формальная структура (термин-слово, термин-словосочетание с указанием компонентов);
- содержательная структура (однозначные, многозначные);
- антропоморфные термины (термины-эпонимы, термины-антропонимы, термины-метафоры);
- нормативные и ненормативные терминологические единицы;
- номенклатурные названия и т. д.

Такая классификация терминов позволит решать ряд прикладных задач, например при исследовании тональности текста можно использовать классификацию нормативной и ненормативной терминологии. Последняя может содержать термины с положительной, нейтральной или негативной коннотацией, например: флюшка/флюра (флюорография) используются с нейтральной коннотацией, а херург вместо хирург обладает ярко выраженной отрицательной коннотацией.

Заключение

В статье описан пример применения метода извлечения русскоязычных многокомпонентных терминологических единиц из научно-технических текстов на основе структурных моделей терминологических единиц. Показано, что термин может быть однокомпонентным и состоять из ключевого слова или представлять собой терминологическое словосочетание, а наиболее продуктивным способом номинации является образование составных терминов, состоящих из двух, трех, четырех и пяти компонентов. Основные структурные модели многокомпонентных терминологических словосочетаний состоят из 12 элементов: имя существительное в именительном падеже, имя существительное в родительном падеже, имя существительное в винительном па-

деже, имя существительное в творительном падеже, имя существительное в предложном падеже, имя прилагательное, имя числительное, наречие, причастие, деепричастие, предлог. Метод извлечения русскоязычных многокомпонентных терминологических единиц из научно-технических текстов на основе структурных моделей терминологических единиц состоит из следующих этапов: анализ предложения по частям речи, исключение частей речи, не входящих в состав терминологических словосочетаний, проверка терминов-кандидатов на стоп-слова, которые образуют широко используемые коллокации с терминами, но в совокупности не являются терминами по семантическому содержанию, проверка терминов-кандидатов по словарю корпуса или передача терминологу на ручную обработку. Перспективой дальнейшего исследования является усовершенствование метода извлечения многокомпонентных терминов путем приписывания дополнительных классификационных тегов к термину, а также исследование возможностей использования описанного метода для других языков.

Список литературы

1. *Захаров В. П., Хохлова М. В.* Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов // Структурная и прикладная лингвистика. 2012. № 9. С. 222–233.
2. *Клышинский Э. С., Кочеткова Н. А., Карпик О. В.* Метод выделения коллокаций с использованием степенного показателя в распределении Ципфа // Новые информационные технологии в автоматизированных системах. 2018. № 21. С. 220–225.
3. *Кочеткова Н. А.* Метод извлечения технических терминов с использованием усовершенствованной меры странности // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 5. С. 25–32.
4. *Кружков М. Г.* Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. 2015. Т. 25. № 2. С. 140–159.
5. *Лейчик В. М.* Оптимальная длина и структура термина // Вопросы языкознания. 1981. № 2. С. 63–73.
6. *Новикова А. А.* Сравнение инструментов Sketch Engine и Termostat для извлечения терминологии // International Journal of Open Information Technologies. 2020. Т. 8. № 11. С. 73–79.
7. *Becerro F. B.* Phraseological variations in medical-pharmaceutical terminology and its applications for English and German into Spanish translations // SciMedicine Journal. 2020. Vol. 2. No. 1. P. 22–29. DOI: 10.28991/SciMedJ-2020-0201-4.
8. *Biziukova N. Yu., Tarasova O. A., Rudik A. V. et al.* Automatic recognition of chemical entity mentions in texts of scientific publications // Automatic Documentation and Mathematical Linguistics. 2020. Vol. 54. No. 6. P. 306–315. DOI: 10.3103/S0005105520060023.
9. *Butenko Iu. I., Garazha V. V.* BMSTU Corpus of scientific and technical texts: conceptual framework // Applied Linguistics Research Journal. 2021. Vol. 5. No. 3. P. 76–81. DOI: 10.14744/alrj.2021.15579.
10. *Loukachevitch N., Dobrov B.* Ontological resources for representing security domain in information-analytical system // Открытые семантические технологии проектирования интеллектуальных систем. 2018. № 8. С. 185–191.
11. *Sidnyaev N. I., Butenko I. I., Bolotova E. E.* Formal grammar theory in recognition methods of unknown objects // Automatic Documentation and Mathematical Linguistics 2020. Vol. 54. No. 4. P. 215–225. DOI: 10.3103/S000510552004007X.
12. *Simon N. I., Kešelj V.* Automatic term extraction in technical domain using part-of-speech and common-word features // Proceedings of the ACM Symposium on Document Engineering. 2018. Article 51. P. 1–4. DOI: 10.1145/3209280.3229100.
13. *Terryn A. R., Hoste V., Lefever E.* In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora // Language Resources and Evaluation. 2020. No. 54. P. 385–418. DOI: 10.1007/s10579-019-09453-9.

Сведения об авторах

Бутенко Юлия Ивановна, ORCID 0000-0002-9776-5709, канд. техн. наук, доцент, кафедра теоретической информатики и компьютерных технологий, Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия, iubutenko@bmstu.ru

Строганов Юрий Владимирович, ORCID 0000-0001-5628-7395, старший преподаватель, кафедра программного обеспечения ЭВМ и информационных технологий, Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия, stroganovyv@bmstu.ru
Сапожков Андрей Михайлович, ORCID 0000-0001-8424-5686, студент, кафедра программного обеспечения ЭВМ и информационных технологий, Московский государственный технический университет им. Н. Э. Баумана, Москва, Россия, andreysapozhkov535@gmail.com

Статья поступила 04.08.2021, рассмотрена 09.08.2021, принята 22.09.2021

References

1. Zakharov V. P., Khokhlova M. V. Automatic term extraction and statistical analysis in a special text corpus as a tool for thesaurus construction. *Strukturnaya i prikladnaya lingvistika*, 2012, no.9, pp.222-233 (in Russian).
2. Klyshinskii E. S., Kochetkova N. A., Karpik O. V. *Metod vydeleniya kollokatsii s ispol'zovaniem stepennogo pokazatelya v raspredelenii Tsipfa* [Collocation extraction method using the power-law index in the Zipf distribution]. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh*, 2018, no.21, pp.220-225.
3. Kochetkova N. A. *Metod izvlecheniya tekhnicheskikh terminov s ispol'zovaniem usovershenstvovannoi mery strannosti* [Method for the extraction of technical terms using an improved measure of strangeness]. *Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnye protsessy i sistemy*, 2015, no.5, pp.25-32.
4. Kruzhkov M. G. Information resources for contrastive studies: electronic text corpora. *Sistemy i sredstva informatiki*=Systems and Means of Informatics, 2015, vol.25, no.2, pp.140-159 (in Russian).
5. Leichik V. M. *Optimal'naya dlina i struktura termina* [Optimal length and structure of the term]. *Voprosy yazykoznaniya*, 1981, no.2, pp.63-73.
6. Novikova A. A. Sketch Engine and Termostat tools for automatic term extraction. International Journal of Open Information Technologies, 2020, vol.8, no.11, pp.73-79 (in Russian).
7. Becerro F. B. Phraseological variations in medical-pharmaceutical terminology and its applications for English and German into Spanish translations. SciMedicine Journal, 2020, vol.2, no.1, pp.22-29. DOI: 10.28991/SciMedJ-2020-0201-4.
8. Biziukova N. Yu., Tarasova O. A., Rudik A. V. et al. Automatic recognition of chemical entity mentions in texts of scientific publications. Automatic Documentation and Mathematical Linguistics, 2020, vol.54, no.6, pp.306-315. DOI: 10.3103/S0005105520060023.
9. Butenko Iu. I., Garazha V. V. BMSTU Corpus of scientific and technical texts: conceptual framework. Applied Linguistics Research Journal, 2021, vol.5, no.3, pp.76-81. DOI: 10.14744/alrj.2021.15579.
10. Loukachevitch N., Dobrov B. Ontological resources for representing security domain in information-analytical system. *Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem*, 2018, no.8, pp.185-191.
11. Sidnyaev N. I., Butenko I. I., Bolotova E. E. Formal grammar theory in recognition methods of unknown objects. Automatic Documentation and Mathematical Linguistics 2020, vol.54, no.4, pp.215-225. DOI: 10.3103/S000510552004007X.
12. Simon N. I., Kešelj V. Automatic term extraction in technical domain using part-of-speech and common-word features. Proceedings of the ACM Symposium on Document Engineering, 2018, article 51, pp.1-4. DOI: 10.1145/3209280.3229100.
13. Terryn A. R., Hoste V., Lefever E. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation, 2020, no.54, pp.385-418. DOI: 10.1007/s10579-019-09453-9.

About the authors

Iuliia I. Butenko, ORCID 0000-0002-9776-5709, Cand. Sci. (Eng.), Associate Professor, Theoretical Informatics and Computer Technologies Department, Bauman Moscow State Technical University, Moscow, Russia, iubutenko@bmstu.ru
Yurii V. Stroganov, ORCID 0000-0001-5628-7395, Senior Lecture, Computer Software and Information Technology Department, Bauman Moscow State Technical University, Moscow, Russia, stroganovyv@bmstu.ru
Andrei M. Sapozhkov, ORCID 0000-0001-8424-5686, Student, Computer Software and Information Technology Department, Bauman Moscow State Technical University, Moscow, Russia, andreysapozhkov535@gmail.com

Received 04.08.2021, reviewed 09.08.2021, accepted 22.09.2021