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-