АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ ТЕРМИНОЛОГИЧЕСКИХ СЛОВОВСОЧЕТАНИЙ

Мусурманкулова Мадина Носировна

Доктор философии (PhD) по педагогическим наукам,

в/и/о доцента кафедры русского языка и литературы,

начальник отдела контроля качества образования

Денауского института предпринимательства и педагогики,

г. Денау, Узбекистан

madinanosirovna87@gmail.com

Аннотфция: *В статье рассматривается проблемы автоматического выделения терминологических словосочетаний. Исследуется с точки зрения многочисленных приложений – создания терминологических словарей на основе корпусов текстов, автоматического индексирования текстов для систем иформационного поиска, рубрикации текстов и их тематической структуризации, перевода текстов с одного языка на другой, извлечения знаний из текстовых источников.*

Ключевые слова: *текст, тематическая структура, источники, словари.*

В настоящей работе проблема автоматического выделения терминов рассматривается в рамках общей задачи создания специализированной системы литературно-научного редактирования научно-технических текстов, [11, С.2], где обсуждаются основополагающие черты такой системы. Особенностью предложенного подхода к автоматизации научного редактирования является учет основных черт научно-технической прозы, прежде всего, ее высокой стандартизированности, ограниченности ее словарного состава, насыщенности терминолической лексикой, как и фразеологическими словосочетаниями общенаучной речи, по большей

части, глагольно-именными [12, С.3].

«К терминологической лексике относятся слова, точно обозначающие определенные понятия какой-нибудь области науки, техники, производства, сельского хозяйства, экономической и общественной жизни, литературы и искусства» [4, с. 32]; «...слово в составе терминологической лексики — это слово, соотносимое со специфическим объектом. Следовательно, терминологическая лексика — это словарный запас, используемый в той или иной отрасли материального производства или науки для обозначения специфических объектов» [13, с. 145].

Как следствие, проблема выделения рассматривается нами более широко. Кроме выделения собственно терминов, имеющих проблемно-ориентированный характер, необходимо обнаружить в тексте терминологизированные словосочетания общенаучной лексики. Конечной целью выделения является не только проверка согласованности употребления терминов и выявление стилистических ошибок в использовании общенаучных слов, но и “свертка” выделенных многословных сочетаний в законченные единицы, что существенно сокращает многовариантность проводимого затем полного синтаксического разбора.

В силу изложенного, выделение разных устойчивых словосочетаний основано на нескольких разработанных для системы научного редактирования компьютерных словарей, вкупе отражающих специфику лексики научно-технической прозы. Для выделения словосочетаний применяется полный морфологический и частичный синтаксический анализ предложений текста. Ниже описываются словарные средства и процедуры, разработанные и используемые для выделения научных терминов и терминологизированных словосочетаний [7, С.1].

При автоматическом выделении словосочетаний используется три словаря: терминологический, словарь сочетаний общенаучной речи и общий морфологический словарь, их связующий.

Морфологический словарь основ и неизменяемых слов покрывает все слова, встречающиеся в первых двух словарях. Кроме грамматической информации (части речи описываемого слова и его флективного класса) словарные статьи содержат отсылки к тем единицам других словарей, которые содержат в своем составе данное слово.

Терминологический словарь системы автоматизированного редактирования научных текстов разработан для проблемной области «Информатика и вычислительная техника», в его основу легли несколько текстовых словарей [7, С.2].

Особенностью устойчивых именных словосочетаний-терминов является их неразрывность, означающая, что большинство их вхождений в текст отличается от зафиксированных в словаре образцов лишь грамматическими окончаниями некоторых слов, составляющих словосочетание. Кроме таких простых вхождений терминов в тексте возможны, хотя и менее часты, сочинительные конструкции, полученные сокращением общего начала или конца нескольких терминов: *ЭВМ второго, третьего и четвертого поколения*, *векторный или растровый дисплей* (при наличии в словаре терминов *ЭВМ второго поколения* , *ЭВМ третьего поколения* и *ЭВМ четвертого поколения*, *векторный дисплей* и *растровый дисплей*).Такие сочинительные сокращения могут быть получены с помощью запятой и союзов (одиночных – *и*, *или*, *либо*, и двойных – *или...или, и...и, не...а, как...так и*), например: *ЭВМ не второго, а третьего поколения* [12, С.2].

Перед собственно выделением терминов каждое предложение текста разбивается на фрагменты по границам всех встреченных специальных знаков, которые не могут встретиться внутри самих терминов и их сочинительных сокращений (например, скобки) и, следовательно, не могут разрывать термины-словосочетания. Далее фрагменты рассматриваются независимо. Сначала делается попытка выделить простые вхождения терминов (максимальные по длине словосочетания, записанные в словаре).

При этом производится поиск во фрагменте непустой и неразрывной цепочки слов, совпадающей со словарным словосочетанием, с точностью до согласованной замены окончаний слов. Выделение сочинительных сокращений производится при повторном просмотре фрагмента, при одновременном движении слева направо и слева направо, позволяющем выявить левые и правые сокращения.

ЛИТЕРАТУРНЫЙ СПИСОК:

1. Bourigault, D. (1992) Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. Proceedings of COLING-92, Nantes, France, p. 977-981.

2. Головин Б. Н. О некоторых проблемах изучения термина.— Вестник Московского ун-та. Серия X. Филология. М., 1972. С.49-59.

3. Dowlagar S., Mamidi R. Unsupervised technical domain terms extraction using term extractor. Proc. XVII ICON, 2020, pp. 5–8.

4. Дементьева Я.Ю., Бручес Е.П., Батура Т.В. Извлечение терминов из текстов научных статей // Программные продукты и системы. 2022. Т. 35. № 4. С. 689–697. DOI: 10.15827/0236235X.140.689-697.689 с.

5. Кобрин Р. Ю. Опыт лингвистического анализа терминологии. Автореф. дис. на соиск. учен, степени канд. филол. наук. Горький, 1969.С. 121.

6. Овчаренко В. М. Концептуальная, семантическая и семиотическая целостность термина. — В кн.: Лингвистические проблемы научно-технической терминологии. М., 1970. С. 139-153.

7. Овчаренко В. М. Термин, аналитическое наименование и номинативное определение. — В кн.: Современные проблемы терминологии в науке и технике. М., 1969.С.91-122.

8. Реформатский А. А. Введение в языкознание. М., 1955. С.536. Smadja, F. (1993) Retrieving Collocations from Text: Xtract. Computational Linguistics, 19 (1), p. 143-177.

9. Smadja, F. (1993) Retrieving Collocations from Text: Xtract. Computational Linguistics, 19 (1), p. 143-177.

10. https://infopedia.su/3xca34.html
11. https://studfile.net/preview/4200604/page:10/
12. https://www.dialog-21.ru/digest/2001/articles/bolshakova/