Комбинирование методов извлечения научных терминов из текстового документа

Семак Владислав Викторович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: vlad.semakk@gmail.com

Научный руководитель *— Большакова Елена Игоревна*

Одной из сложных задач в области обработки текстов на естественном языке является терминологический анализ проблемно-ориентированных текстов, предполагающий автоматическое выделение набора терминов, а также ключевых слов, которые отражают основной смысл текстового документа. Данный вид анализа текста используется на практике для создания глоссариев, т.е. перечней терминов с их определениями, и предметных указателей [1], содержащих значимые термины текста с указанием номеров страниц, облегчающих поиск нужной информации.

В данной работе рассматриваются методы автоматического извлечения терминов из отдельного научного текста на русском языке. Сложность решаемой задачи связана с тем, что большинство статистических методов выявления терминов разработано для текстовых коллекций, для отдельных текстов эти методы работают значительно хуже. С целью повышения эффективности извлечения терминов предлагается применить комбинацию несколько известных подходов: грамматические шаблоны для извлечения однословных и многословных терминов, статистические меры для ранжирования извлеченных терминов-кандидатов, их переранжирование на основе графов совместной встречаемости слов термина.

Для извлечения и ранжирования терминов была реализована и экспериментально исследована комбинация следующих методов:

- извлечение терминов-кандидатов на основе грамматических шаблонов, описывающих часть речи слов термина, а также типичные фразы определения терминов в тексте;
- фильтрация извлеченных кандидатов на основе частоты их встречаемости в тексте и списка стоп-слов (слов, которые не могут быть терминами или их частью);
- ранжирование с использованием статистической меры терминологичности C-value и разновидности алгоритма графового ранжирования [2].

Проверка эффективности реализованной комбинации проводилась на русскоязычных учебно-научных текстах: учебных пособиях по дискретной математике, функциональному программированию, искусственному интеллекту и формальным грамматикам. Для каждого обработанного текста был автоматически получен ранжированный список извлеченных терминов-кандидатов, релевантность которых оценивалась экспертами. Эксперименты показали, что предлагаемая комбинация позволяет достичь качества до 78% средней точности, что значительно превышает качество каждого из объединяемых методов.

Литература

1. Bolshakova E. I., Ivanov K. M. Automating Hierarchical Subject Index Construction for Scientific Documents.// In: The Eighteenth Russian Conference on Artificial Intelligence RCAI-2020, Lecture Notes on Artificial Intelligence, Vol. 12412, Springer, 2020, pp. 201-214.
2. Zhang Z., Gao J., Ciravegna F. SemRe-Rank: Improving Automatic Term Extraction by Incorporating Semantic Relatedness with Personalised PageRank// Association for Computing Machinery, Vol. 12, New York, NY, USA, 2018.