Combining methods for term extraction from scientific and technical text

Е. И. Большакова[1] В. В. Семак[2]

В докладе рассматривается подход к автоматическому извлечению терминов из научно-технического текста, комбинирующий известные методы: лингвистические шаблоны, статистические меры терминологичности, методы графового ранжирования. Описываются комбинируемые методы и этапы для извлечения, отбора и ранжирования терминов, реализованные для обработки документов на русском языке. Приводятся результаты экспериментов по извлечению терминов из учебных текстов по математике и программированию. Полученные оценки эффективности извлечения (74% средней точности) показывают перспективность описанного подхода.

Ключевые слова: обработка текстов на естественном языке, автоматическое извлечение терминов, лингвистические шаблоны, графовые методы ранжирования

1. Введение

Одной из задач автоматической обработки текстов на естественном языке является терминологический анализ специализированных текстов, предполагающий извлечение из текстов терминов, выражающих понятия предметной области. Большинство известных методов извлечения терминов разработано для анализа текстовых коллекций, с целью построения терминологических словарей, тезаурусов, онтологий предметных областей. Однако, как показано в недавней статье [3], при работе с отдельно взятым текстом эти методы, преимущественно статистические, обычно показывают худшие результаты (порядка 23-65% средней точности и 5-38% F-меры). Тем не менее извлечение терминов из заданного текстового документа, особенно узкоспециализированного (научной статьи, книги, учебника, технического руководства и т.д.), необходимо на практике для автоматизации аннотирования и реферирования текстов,

[1] *Большакова Елена Игоревна* — доцент, к.ф-м.н каф. алгоритмических языков ВМК МГУ, e-mail: eibolshakova@gmail.

Bolshakova E. I. — Ph.D, Senior Lecturer, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Algorithmic Language Department.

[2] *Семак Владислав Викторович* — аспирант каф. алгоритмических языков ВМК МГУ, e-mail: vlad.semakk@gmail.com.

Semak V. V. — graduate student, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Algorithmic Language Department.

создания глоссариев (перечней терминов с их определениями) и предметных указателей (обратных индексов к основным терминам текста с указанием номеров их страниц).

В данной работе рассматривается задача автоматического извлечения однословных и многословных терминов из отдельного русскоязычного научно-технического текста. Для повышения эффективности ее решения предлагается применить комбинации нескольких известных методов: лингвистические шаблоны терминов и контекстов их употребления в научных текстах, статистические меры для ранжирования извлеченных терминов-кандидатов, их переранжирование на основе графов совместной встречаемости слов. Рассмотренные комбинации методов экспериментально исследованы на учебно-научных текстах среднего объема, получены оценки их эффективности по мере средней точности (average precision), обычно используемой для задач извлечения терминов.

2. Извлечение и ранжирование терминов

Исследуемые в данной работе комбинации методов воплощены в трехэтапной стратегии выявления значимых терминов текста.

На первом этапе выполняется анализ исходного текста и извлечение списка терминов-кандидатов (слов и словосочетаний), удовлетворяющих лексико-синтаксическим шаблонам нескольких видов. Шаблоны описывают грамматические образцы терминов (часть речи и другие грамматические характеристики входящих в них слов), а также формализуют контексты определений терминов и их синонимов, типичные для научных текстов.

На втором этапе происходит фильтрация извлеченных кандидатов в термины с помощью заранее составленных списков стоп-слов (слов, которые не могут быть терминами или их частью), а затем из оставшегося множества производится последовательный отбор терминов на основе нескольких факторов [1]:

- достоверности шаблона, которым был извлечен термин-кандидат (шаблоны имеют разную точность распознавания терминов), и в первую очередь в результирующее множество отбираются кандидаты, выделенные наиболее достоверными шаблонами;

- частоте встречаемости термина-кандидата в тексте: согласно закону Ципфа значимые термины должны принадлежать центральной части распределения кандидатов по частоте встречаемости в тексте (не могут быть как слишком частотными, так и слишком редкими);

- лексической схожести терминов одной тематики: в результирующее множество добавляются кандидаты, имеющие общие слова с уже отобранными терминами (например: *сходимость функционального ряда и функциональный ряд*).

На третьем этапе происходит ранжирование отобранного множества терминов- кандидатов с целью их упорядочивания по релевантности предметной области. Рассматривались три способа ранжирования терминов-кандидатов: широко используемая статистическая мера терминологичности C-value [2], метод графового ранжирования – персонализированный PageRank [4], а также их комбинация. Результатом решения задачи автоматического извлечения терминов считаются первые 90% элементов ранжированного списка, граница отсечения для ранжированного списка подбиралась экспериментально.

3. Результаты экспериментов

Проверка эффективности реализованной стратегии извлечения проводилась на текстах семи русскоязычных учебно-научных текстов среднего размера (12-55 тыс. слов) по темам: формальные грамматики (ФГ), дифференциальные уравнения (ДУ), дискретная математика (ДМ), искусственный интеллект (ИИ), язык программирования Лисп (ЯЛ), системы программирования (СП) и математический анализ (МА). Для каждого обработанного текста был автоматически получен ранжированный список извлеченных терминов, их релевантность оценивалась экспертами.

В Таблице 1 приведены оценки средней точности (average precision) извлечения для трех вариантов ранжирования: персонализированный PageRank, C-value и их комбинация. Лучшего качества на четырёх текстах (ФГ, ДУ, ЯЛ, МА) достигает комбинация C-value и PageRank, на трёх остальных текстах качество не уступает или незначительно (менее 1%) уступает средней точности, полученной с использованием C-value. Итоговое ранжирование на основе PageRank не дает преимуществ. Таким образом, ранжирование терминов с использованием комбинации C-value и персонализированного PageRank или одного C-value может использоваться практически равноценно.

Нижняя строка Таблицы 1 (средняя оценка по всем обработанным текстам) показывает, что описанный в работе подход с применением для ранжирования меры C-value или ее комбинации с PageRank, позволяет достичь большей эффективности (в среднем 74% и даже до 88% для текста по ФГ), чем отдельные методы, исследованные в [3].

Таблица 1. Средняя точность извлечения терминов

Текст	PageRank	C-value	C-value+PageRank
ФГ	0.80	0.87	**0.88**
ДУ	0.61	0.74	**0.75**
ДМ	0.57	**0.67**	**0.67**
ИИ	0.43	**0.73**	0.72
ЯЛ	0.75	0.79	**0.80**
СП	0.44	**0.64**	0.63
МА	0.65	0.71	**0.72**
Среднее	0.61	0.73	**0.74**

Список литературы

[1] Bolshakova E. I., Ivanov K. M., "Automating Hierarchical SubjectIndex Construction for Scientific Documents", *The EighteenthRussian Conference on Artificial Intelligence RCAI-2020, LectureNotes on Artificial Intelligence, Springer*, **12412** (2020), 201–214.

[2] Frantzi K., Ananiadou S., Mima H., "Automatic Recognition of Multi-Word Terms: The C-value/NC-value method", *International Journal on Digital Libraries*, **3**:2 (2000), 115–130.

[3] Šajatović A. et al., "Evaluating automatic term extraction methods on individual documents", *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 2019, 149–154.

[4] Zhang Z., Gao J., Ciravegna F., "SemRe-Rank: Improving AutomaticTerm Extraction by Incorporating Semantic Relatedness with Personalised PageRank", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **12**:5 (2018), 1–41.

E.I. Bolshakova, V.V. Semak

Комбинирование методов для извлечения терминов из научно-технического текста

An approach to automatic extraction of terms from an individual scientific text is reported, which combines known methods: linguistic patterns, statistical terminological measures, methods of graph ranking. The combined methods and stages for extracting, selection and ranking of terms are described, which are implemented for processing documents in Russian. The results of experiments on extracting terms from educational texts in mathematics and programming are presented. The scores of extraction efficiency (74% of average accuracy) show that the described approach is promising.

Keywords: natural language processing, automatic term extraction, linguistic templates, graph ranking methods.