ФИЛОЛОГИЯ. ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА
PHILOLOGY. THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

Научная статья *Филологические науки*
УДК 81’33 *https://doi.org/10.26907/2658-3321.2023.6.3.388-396*

МЕТОДЫ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ В НАУЧНЫХ ТЕКСТАХ (НА МАТЕРИАЛЕ СТАТЕЙ ПО НАПРАВЛЕНИЮ НАУКИ О ЗЕМЛЕ)

Т.С. Падерина
Иркутский научный центр Сибирского отделения Российской академии наук, Иркутск, Россия
jana-pad@mail.ru, https://orcid.org/0000-0002-2603-6242

Аннотация. Статья посвящена описанию теоретических и прикладных положений первоначального этапа работы по автоматическому извлечению терминов из научных текстов. Данный этап работы является частью государственного задания научной лаборатории лингво-педагогических исследований по теме «Лингвосемиотическая гетерогенность научной картины мира: теоретическое и лингводидактическое описание». Цель исследования заключается в извлечении терминов из подготовленного корпуса научных текстов, относящихся к определенной предметной области. Для этого был использован корпус научных текстов по направлению Науки о Земле, подготовленный методом случайной выборки при помощи приложения Semantic Scholar. Извлечение терминов при помощи автоматической обработки текстов (АОТ) является перспективным направлением исследования, так как позволяет упростить процесс создания терминосистем или составления онтологии для узкоспециализированных предметных областей. В условиях быстро меняющегося потока информации данный вид работы с текстами, безусловно остается актуальным направлением и позволяет быстрее и эффективнее обрабатывать большие объемы материалов. Однако, необходимо отметить, что автоматическое извлечение терминов (АОТ) не всегда является точным и может содержать ошибки. Поэтому, важно проводить дополнительную проверку и корректировку полученных результатов. Перспективы исследования связаны с совершенствованием существующих инструментов автоматической обработки текстов (АОТ). Кроме этого, анализ извлеченных терминов позволил нам сформировать основу для дальнейших практических исследований по созданию цифрового продукта (цифровой модели определенных терминосистем) для хранения, систематизации и использования терминосистем по определённой узкоспециализированной предметной области.

Ключевые слова: терминология; извлечение терминов; тематическое моделирование; научная коммуникация

Для цитирования: Падерина Т.С. Методы извлечения терминов в научных текстах (на материале статей по направлению науки о земле). *Казанский лингвистический журнал*. 2023;6(3): 388–396. *https://doi.org/10.26907/2658-3321.2023.6.3.388-396*

Original article *Philology studies*
https://doi.org/10.26907/2658-3321.2023.6.3.388-396

METHODS FOR TERMINOLOGY EXTRACTION IN SCIENTIFIC TEXTS (BASED ON ARTICLES OF EARTH SCIENCES)

T.S. Paderina

Irkutsk Scientific Center of Siberian Branch of Russian Academy of Sciences, Irkutsk, Russia
jana-pad@mail.ru, https://orcid.org/0000-0002-2603-6242

Abstract. The article describes the theoretical and applied provisions of the initial stage of work on automatic extraction of terms from scientific texts. This stage of the work is a part of the state assignment of the Scientific Laboratory of Linguistic and Pedagogical Research on "Linguosemiotic heterogeneity of scientific picture of the world: theoretical and linguodidactic description". The aim of the research is to extract terms from a prepared corpus of scientific texts relating to a particular subject area. For this purpose, a corpus of scientific texts in the field of Earth Sciences, prepared by random sampling using the Semantic Scholar application, was used. The term extraction by automatic text processing (ATP) is a promising area of research as it simplifies the process of creating terminology systems or ontologies for highly specialized subject areas. With the rapidly changing flow of information, this type of work with texts is undoubtedly still relevant and allows for faster and more efficient processing of large volumes of material. However, it should be noted that automatic term extraction is not always accurate and may contain some errors. Therefore, it is important to carry out additional verification and correction of the results obtained. Prospects for the study are related to the improvement of existing automatic text processing tools. In addition, the analysis of the extracted terms has enabled us to form the basis for further practical research into the creation of a digital product (a digital model of certain terminology systems) for the storage, systematization and use of terminology systems for a certain highly specialized subject area.

Keywords: terminology; terminology extraction; thematic modeling; scientific communication

For citation: Paderina T.S. Methods for Terminology Extraction in Scientific Texts (Based on Articles of Earth Sciences). *Kazan Linguistic Journal*. 2023;6(3): 388–396. (In Russ.). *https://doi.org/10.26907/2658-3321.2023.6.3.388-396*

Стремительное развитие науки, применение цифровых технологий для работы с большим объемов данных определяет актуальность разработки и совершенствования методов сбора, хранения и обработки информации при помощи автоматических программ. В нашей работе мы рассматриваем интерфейсы, которые могут оказать помощь исследователям при работе с текстами научных статей (поиск статей соответствующего профиля, беглый просмотр для достижения понимания содержания, поиск ключевых слов, интеграция знаний, полученных при анализе статей, в свои исследовательские работы.

Принцип отбора тестов обусловлен задачами Государственного задания по теме «Лингвосемиотическая гетерогенность научной картины мира: теоретическое и лингводидактическое описание» (0275-2022-0001). Тексты должны быть научным (написанным с соблюдением норм академического дискурса), быть написанным на английском языке, кроме этого, мы добавили дополнительные параметры, а именно: текст должен быть не старше 5 лет и входить в

журналы группы Scopus (Q1) или Web of Science (WoS). Такой выбор позволяет допустить к выборке контексты, максимально соответствующие правилам письменной научной коммуникации.

Работа с научными текстами (как в прочем и с текстами других дискурсов, такими как художественный, политический и т.д.) для статистической работы и анализа проще рассматривать в синхронии, чем в диахронии, т.к. такой подход позволяет задействовать меньший объем текстов для выявление определенных закономерностей и не рассматривать такие факторы как исторические, социальные, политические и иные трансформации, которые оказывали влияние на формирование терминосистем. Изучение терминосистем в диахронии на данный момент недостаточно изученная тема. Вопрос о том, какие движущие силы стоят за терминологизацией, специализацией или лексическими изменениями в ограниченных областях, насколько нам известно, даже не затрагивался.

В нашей работе мы предприняли попытку описать результат анализа по извлечению терминологии в англоязычных научных текстах. В качестве материала для извлечения терминов был использован корпус статей за 2015–2022 гг. по направлению подготовки «Науки о Земле» из журналов *Landslides, Natural Hazards and Earth System Sciences, Journal of Geophysical research: Earth Surface, Geophysical Research Letters* и другие.

Анализируя терминологию в подмножестве текстов, отобранных методом случайной выборки через ключевые слова, мы надеемся сформировать основу для дальнейших эмпирических исследований узкопрофильной терминологии и создания программного продукта с целью хранения, систематизации и доступа к определённой терминосистеме в автоматическом режиме. Другими словами, цель работы – выявление цифровой модели для извлечения терминов в научных текстах на английском языке по определённому направлению подготовки и создание программного продукта (цифровой модели терминосистемы) с возможностью дальнейшего её совершенствования и автоматического пополнения.

Прежде чем перейти к решению задачи по распознаванию информации из текстов мы осуществили обзор существующих методов извлечения терминов, который показал три основных подхода:

1. Извлечение терминов на основании статистических показателей (вероятностный метод) [1]. Суть метода основывается на данных частотности встречаемости словосочетаний. То есть выделяется две или более лексические единицы, частота совместной встречаемости которых выше первоначально заданного уровня. Слабая сторона данного метода, которую мы отмечаем, – длина словосочетания, так как при увеличении длины термина, его частотность уменьшается.

2. Извлечение терминов на основе правил (rule-based). В основе заранее подготовленный исследователями свод правил (лингвистические шаблоны). Данный метод применяется с опорой на словари и онтологии, дает хорошую точность, но только для конкретного языка (в силу особенностей грамматики заданных языков).

3. Извлечение терминов при помощи алгоритмов машинного обучения (machine learning). Извлечение цепочек слов, которые могут быть терминами и дальнейшее определение границ данного термина. (тексты делятся на заданные интервалы или на интервалы с наибольшей частотностью совпадений). В рамках этого подхода применяются такие методы обучения с учителем (supervised), методы обучения без учителя (unsupervised), методы частичного обучения с учителем (bootstrapping). Чаще всего выбор останавливается на методе обучения с учителем, который подразумевает построение программной модели (машинный классификатор), способной отличать искомые данные от всех остальных. Обучение модели (построение машинного классификатора) осуществляется на текстах, размеченных вручную, в которых значимым объектам прописывают определенные метки (теги).

После анализа подходов было принято решение об использовании комбинированного метода с применением положительных сторон каждого из подхо-

дов. Одной из задач автоматической обработки научных текстов является анализ терминологии, выражающей понятия определенной предметной области. Извлечение терминов из узкоспециализированных текстов необходимо для автоматизации аннотирования и реферирования текстов, создания глоссариев и т.д. в нашем исследовании рассматривается задача автоматического извлечения однословных и многословных терминов из специализированного научного текста на английском языке. Извлечение терминов выполнялось в 3 этапа:

1 этап. Анализ научного текста и извлечение списка терминов-кандидатов (слова, словосочетания), которые будут соответствовать лексико-синтаксическим шаблонам, отображающие характерные конструкции научных текстов (N+N, N+A, A+N, где N – существительное, A – прилагательное и т.д.);

2 этап. Фильтрация извлеченных терминов-кандидатов с помощью определенного списка стоп-слов (stop-word-list) (слова, которые не могут быть частью термина, например “*is a*” или “*of the*”). Объектом внимания становятся более сложные структур, такие как словосочетания, комбинация существительных и т.д. мы предполагаем, что термины, имеющие сложную структуру, должны быть составлены из терминов, обладающих изначально простой структурой.

3 этап. Упорядочивание терминов-кандидатов по релевантности предметной области. На данном этапе исследования мы производим ранжирование используя статистическую меру терминологичности C-value [2]:

$$C\text{-}value(a) = (length\ (a) - 1)\ (n(a) - \frac{t(a)}{c(a)}$$

где *(a)* – сложное существительное (compound noun), *length (a)* – количество однокоренных слов, составляющих (a), *n(a)* – общая частота встречаемости (a)в корпусе текстов, *t(a)* – частота встречаемости в более длинных терминах-кандидатах и c*(a)* – общее количество этих терминов.

Несмотря на кажущееся разнообразие существующих программных цифровых продуктов для автоматической работы с текстами (первые прикладные исследования по извлечению информации из специализированных текстов от-

 ISSN: 2658-3321 (Print)

носятся к началу 80-х годов XX века [3]) задача извлечения терминов из научных текстов остается актуальной. При работе по извлечению терминов из научных текстов мы столкнулись со следующими проблемами:

1. Определение границ сложносоставных терминов.
2. Распознавание лексических единиц как части составного термина.
3. Определение лексических единиц как термина в зависимости от контекста.

Поскольку работа с большим объемов текстов вручную достаточно сложный процесс, за последнее время было предложено большое количество программных продуктов, позволяющих автоматизировать данный процесс [4, 5,6,7]. В нашей исследовательской работе мы обратились к программе MALLET (*англ.* MAchine Learning for LanguagE Toolkit). MALLET – это инструмент тематического моделирования полезный для статистической обработки текстов естественного языка, для классификации документов, разметки последовательностей, кластерного анализа, извлечения информации, и т.д. Тематическое моделирование послужило способом для способ анализа больших объемов немаркированного текста. Мы задали «тему», группы слов, которые часто встречаются вместе (термины, терминологические словосочетания) и, используя контекстуальные подсказки, тематическая модель (thematic model) соединила слова с похожими значениями, кроме того, она позволила различить использование слов с несколькими значениями. Следует отметить, что тематическое моделирование позволило нам решить ряд задач:

- выявить смысл или тематику документа по их содержанию
- осуществить классификацию документов по заданным параметрам
- осуществить индексацию и автоматическое аннотирование и т.д.

Из очевидных минусов на данном этапе исследования мы отмечаем, что фактически из текстов извлекается довольно большой список слов-кандидатов, которые в дальнейшем всё равно должны быть проанализированы и подтверждены экспертом по предметной области.

Особое внимание при исследовании было уделено также отбору текстов относительно небольшого объема, поскольку тестовая аннотация показала, что длина текста оказывает негативное влияние на согласованность терминов. Еще одно ограничение заключалось в том, что выборки, взятые за определенный период времени, должны быть сопоставимы по размеру. В таблице 1 показано, сколько текстов и словосочетаний было выбрано для каждого периода времени. Также указано количество извлеченных терминов.

Таблица 1.

Период	2017	2018	2019	2020	2021	2022
Термины	896	987	1400	585	860	1470
Тексты	7	7	10	9	10	10

После подготовки корпуса текстов мы использовали приложение WebAnno и серию скриптов на Python для манипулирования данными. Чтобы внимательно отследить качество извлечений терминов, были проведены встречи для обсуждения конфликтов. Точнее, мы имели дело с двумя типами конфликтов:

1. Несоответствие идентификации термина, т. е. были отмечены разные, непересекающиеся строки словосочетаний;

2. Несоответствие определения диапазона терминов: разные, но перекрывающиеся строки словосочетаний.

Таким образом мы пришли к следующим промежуточным выводам: для автоматического извлечения терминов из узкоспециализированных текстов применяются методы, основанные на лингвистических и статистических критериях. Статистические данные позволяют продемонстрировать частотность встречаемости слов/ словосочетаний в обрабатываемом тексте (корпусе текстов). Лингвистические критерии опираются на структуру терминов. Структура терминов, как правило описывается в виде шаблона (N+N, N+A, A+N, N+ N_{gen}). По мере того, как обработка естественного языка продвигается к пониманию

 ISSN: 2658-3321 (Print)

естественного языка, то есть к способности машины обрабатывать значение языка, а не только его структуру, растет потребность в корпусах, анализируемых на семантическом уровне. Семантические структуры предъявляют иные требования к инструментам аннотирования, чем морфологические или синтаксические аннотации.

В данной работе были описаны методы извлечения терминов из научных текстов узкой направленности на английском языке. Анализ показал, что модели способны к обобщению, а создание программного продукта может обновляться и пополняться в зависимости от поставленных исследователем задач.

Список литературы

1. Дементьева Я.Ю., Бручес Е.П., Батура Т.В. Извлечение терминов из текстов научных статей. *Программные продукты и системы/Software & Systems.* 2022;35(4):689–697. DOI: 10.15827/0236-235X.140.689-697

2. Большакова Е.И., Семак В.В. Комбинирование методов для извлечения терминов из научно-технического текста. *Интеллектуальные системы. Теория и приложения*. 2021;25(4):239–242.

3. Grishman R. *Information Extraction*. In: The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds). WileyBlackwell; 2010. Pp. 515–530.

4. Бручес Е. П., Батура Т. В. Метод автоматического извлечения терминов из научных статей на основе слабо контролируемого обучения. *Вестник НГУ. Серия: Информационные технологии.* 2021;19(2):5–16. DOI 10.25205/1818-7900-2021-19-2-5-16

5. Рогачева В. Э. Методы извлечения терминологических единиц из корпуса сопоставимых текстов. *Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация.* 2017;(2):118–122.

6. Eckart de Castilho R., Mújdricza-Maydt, É.,et al. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *In Proceedings of the LT4DH workshop at COLING.* 2016. Osaka, Japan.

7. Шейко А.М. Инструменты прикладной лингвистики в контроле качества перевода. *Казанский лингвистический журнал*. 2023;6(2):282–293. DOI 10.26907/2658-3321.2023.6.2.282-293.

References

1. Dement`eva Ya.Yu., Bruches E.P., Batura T.V. Terms extraction from texts of scientific papers. *Programmny`e produkty` i sistemy`/Software & Systems*. 2022;35(4):689–697. DOI: 10.15827/0236-235X.140.689-697 (In Russ.)

2. Bol`shakova E.I., Semak V.V. Combining methods to extract terms from scientific and technical text. *Intellektual`ny`e sistemy`. Teoriya i prilozheniya*. 2021;25(4):239–242. (In Russ.)

3. Grishman R. *Information Extraction*. The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds). WileyBlackwell; 2010. Pp. 515–530.

4. Bruches E. P., Batura T. V. Method for Automatic Term Extraction from Scientific Articles Based on Weak Supervision. *Vestnik NGU. Seriya: Informacionny`e texnologii*. 2021;19(2):5–16. DOI 10.25205/1818-7900-2021-19-2-5-16 (In Russ.)

5. Rogacheva, V. E`. Methods of extracting terminological units from the corpus of comparable texts. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul`turnaya kommunikaciya*. 2017;(2):118–122. (In Russ.)

6. Eckart de Castilho R., Mújdricza-Maydt, É.,et al. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *In Proceedings of the LT4DH workshop at COLING*. 2016. Osaka, Japan (In Eng.)

7. Sheiko A.M. Language technology toolsin translation quality assurance. Kazan Linguistic Journal. 2023;6(2):282–293. DOI 10.26907/2658-3321.2023.6.2.282-293. (In Russ.)

Автор публикации

Падерина Татьяна Сергеевна –
Младший научный сотрудник
Иркутский научный центр Сибирского отделения Российской академии наук
Иркутск, Россия
Email: jana-pad@mail.ru
https://orcid.org/0000-0002-2603-6242

Раскрытие информации о конфликте интересов

Автор заявляет об отсутствии конфликта интересов.

Информация о статье

Поступила в редакцию: 8.06.2023
Одобрена после рецензирования: 10.07.2023
Принята к публикации: 15.07.2023

Автор прочитал и одобрил окончательный вариант рукописи.

Информация о рецензировании

«Казанский лингвистический журнал» благодарит анонимного рецензента (рецензентов) за их вклад в рецензирование этой работы.

Author of the publication

Paderina Tatiana Sergeevna –
Junior Researcher
Irkutsk Scientific Center of Siberian Branch of Russian Academy of Sciences
Irkutsk, Russia
Email: jana-pad@mail.ru
https://orcid.org/0000-0002-2603-6242

Conflicts of Interest Disclosure

The author declares that there is no conflict of interest.

Article info

Submitted: 8.06.2023
Approved after peer reviewing: 10.07.2023
Accepted for publication: 15.07.2023

The author has read and approved the final manuscript.

Peer review info

Kazan Linguistic Journal thanks the anonymous reviewer(s) for their contribution to the peer review of this work.