УДК 81`33, 373

С. И. Монахов[1], В. В. Турчаненко[2], Е. А. Федюкова[3], Д. Н. Чердаков[4]
[1]sergomon@gmail.com
Российский государственный педагогический университет им. А. И. Герцена, Санкт-Петербург, Россия
[2]vladimir.turchanenko@mail.ru
Российский государственный педагогический университет им. А. И. Герцена; Институт русской литературы (Пушкинский Дом) РАН, Санкт-Петербург, Россия
[3]katefedyukova@gmail.com
ООО «ЦРТ», Санкт-Петербург, Россия
[4]dm.cherdakov@gmail.com
Российский государственный педагогический университет им. А. И. Герцена; Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

АНАЛИЗ ТЕРМИНОЛОГИИ В СОВРЕМЕННЫХ ШКОЛЬНЫХ УЧЕБНИКАХ МЕТОДАМИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ *

В статье изложены ход, предварительные результаты и перспективы исследования состава и функционирования терминологической лексики в школьных учебниках с помощью методов автоматического извлечения терминов из созданного полнотекстового корпуса (212 учебников 5–11-х классов, 21 дисциплина), а также с помощью моделей анализа семантики естественных языков Word2Vec и нейронных сетей.

Ключевые слова: термин, терминология, векторное представление, учебник, общее образование, нейросеть, глубокое обучение, Word2Vec.

Sergei I. Monakhov[1], Vladimir V. Turchanenko[2], Ekaterina A. Fedyukova[3], Dmitrii N. Cherdakov[4]
[1]sergomon@gmail.com
Herzen State Pedagogical University of Russia, Saint Petersburg, Russia
[2]vladimir.turchanenko@mail.ru
Herzen State Pedagogical University of Russia;
Institute of Russian Literature Russian Academy of Science, Saint Petersburg, Russia
[3]katefedyukova@gmail.com
Speech technology center, Saint Petersburg, Russia
[4]dm.cherdakov@gmail.com
Herzen State Pedagogical University of Russia; Saint Petersburg State University
Saint Petersburg, Russia

WHAT COMPUTATIONAL LINGUISTICS TECHNIQUES REVEAL ABOUT THE USE OF TERMINOLOGY IN MODERN RUSSIAN TEXTBOOKS

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14032 мк «Изучение терминологических подсистем современных школьных учебников на русском языке с помощью моделей анализа семантики естественных языков Word2Vec и нейронных сетей».

The article describes the progress, preliminary results and the prospects of studying the terminological inventory of modern Russian textbooks by means of automatic extraction of terms from the specially created full-text corpus (212 textbooks, 21 disciplines, 5–11 grades) and analysis of their distributional semantics with the help of Word2Vec neural networks.

Keywords: term, terminology, vector representation, school textbook, general education, neural network, deep learning, Word2Vec.

Введение. Школьная учебная литература в аспекте терминоведения изучена слабо, хотя количество терминологических единиц в школьных учебниках весьма велико. Например, согласно нормативным документам, только по предмету «Русский язык» в 5–11-х классах учащийся должен усвоить значение свыше 1 000 терминов и терминологических сочетаний. Школьные терминологические словари не отражают всего набора терминов, используемых в школьных учебниках, и не раскрывают особенностей их функционирования в школьном учебном тексте. Терминология школьных учебников по разным предметам не изучена в сопоставительном аспекте, например в отношении терминологической плотности учебных текстов по разным дисциплинам. Особенности состава и функционирования школьной терминологии в сравнении с собственно научным терминологическим аппаратом также не освещены должным образом. Традиционные, «ручные» способы извлечения терминов из школьного учебного текста и формирования их перечней не могут обеспечить охват значительных массивов данных и не выявляют частотности употребления терминов, как и их сочетаемостных связей, на основе которых формируются системные отношения между терминами.

В связи с этим перспективной представляется автоматизация вычленения терминов из текстов – актуальная задача машинной обработки естественного языка. Современные системы автоматического извлечения терминов из текстов, как правило, используют статистический подход и опираются на сравнение частоты встречаемости лексических элементов в двух корпусах – целевом (откуда необходимо вычленить термины) и референционном (представляющим систему языка в целом; чаще всего это национальный корпус соответствующего языка). Единицы целевого корпуса, отличающиеся неожиданно высокой частотностью по сравнению с частотностью в референционном корпусе, получают в системе статистическую меру ключевого слова (keyness score) и при превышении данной мерой определенного порога определяются в качестве кандидатов на присвоение им терминологического статуса [1].

Один из недостатков этого подхода заключается в том, что полученный список терминов не отражает системных связей между ними. Возможное решение этой проблемы мы видим в применении к описанному материалу методов дистрибутивной семантики, реализованных в наборе компьютерных алгоритмов Word2Vec (continuous-bag-of-words – CBOW, skip-gram), которые реализуют идею выведения значения слова из его лек-

сического окружения: слова могут быть автоматически сгруппированы по степени семантической близости на основе контекстов, в которых они встречаются [2–5]. Алгоритмы Word2Vec, использующие для анализа семантики естественного языка векторные представления слов, в настоящее время широко востребованы, но, насколько нам известно, еще не использовались для исследования и моделирования терминологических подсистем, в т. ч. в отношении учебных текстов. Представляется, однако, что данные методы могут быть успешно применены к анализу лексического состава школьных учебников в целях разноаспектной стратификации их терминологического наполнения. В частности, становится возможным оценить частотность терминов, терминологическую плотность текста, типичное и нетипичное лексическое окружение термина, синтагматические связи терминов и нетерминов, терминологическое и нетерминологическое употребление одних и тех же слов и др. Важно, что все эти аспекты могут выступать основой сопоставительного анализа учебников по разным предметам, а также учебников для разных классов (от 5-го к 11-му) внутри одной дисциплины.

Ход исследования и промежуточные результаты. Исследование проводилось в несколько этапов. Первой задачей было создание целевого (исследовательского) корпуса текстов современных школьных учебников на русском языке. Был составлен репрезентативный список учебников (212 учебников с 5-го по 11-й класс по 21 дисциплине – алгебре, астрономии, биологии, всеобщей истории и истории России, географии, геометрии, естествознании, изобразительному искусству, информатике, литературе, математике, математическому анализу, мировой художественной культуре, музыке, обществознанию, праву, русскому языку, технологии, физике, физической культуре, химии) из числа входящих в федеральный перечень изданий, рекомендованных Министерством просвещения; было получено согласие издательства «Просвещение» на использование текстов этих учебников в исследовательских целях. После сканирования и распознавания тексты прошли предварительную обработку (удаление небуквенных символов, знаков препинания и др.), затем была осуществлена автоматическая лемматизация словоформ и POS-тэгирование (частеречная разметка). Корпус был разделен на подкорпусы согласно учебным дисциплинам и годам обучения и загружен на платформу Sketch Engine (https://www.sketchengine.eu). По требованиям правообладателя корпус доступен только исследовательской группе. Общий объем целевого корпуса – около 14 370 000 слов. В качестве референционного корпуса был избран Russian Web 2011 Sample (ruTenTen11), доступный в Sketch Engine и содержащий более 900 миллионов слов из русскоязычных интернет-текстов.

Следующим этапом работы стало автоматическое извлечение кандидатов в термины из целевого корпуса согласно описанной выше процедуре сравнения относительной частоты лексем целевого корпуса с относительной частотой аналогичных лексем референционного корпуса.

При извлечении однословных кандидатов в термины (keywords в нотации Sketch Engine) сразу высчитывалась метрика keyness score, для неоднословных сочетаний – кандидатов (terms в нотации Sketch Engine) этой процедуре предшествовал этап вычисления специальной метрики Log-Dice score (подробнее см. [6]). Списки однословных и неоднословных кандидатов в термины были упорядочены по убыванию значения метрики keyness score; первые 1 000 позиций в обоих списках были сохранены для дальнейшей работы.

Главной трудностью, с которой пришлось столкнуться после выполнения описанных процедур, является разграничение собственно терминологической лексики и лексики нетерминологической, но уподобленной терминам по поведению в текстах учебников, то есть низкочастотной в референционном корпусе, но высокочастотной в целевом корпусе (ниже эти слова будут условно именоваться лжетерминами). Например, автоматическое извлечение искомых единиц из подкорпуса учебников по русскому языку приводит к вычленению, помимо лингвистических терминов, лжетерминов, тематически связанных с описанием природы: «роща», «туча», «ландыш», «сумрак», «груша», «овраг» и т. п. Для преодоления указанной трудности была проведена векторизация корпуса, для каждой из учебных дисциплин были созданы и обучены дистрибутивно-семантические модели (word embedding models), позволяющие выявить относительную семантическую близость единиц изучаемых терминологических подсистем. При обучении моделей использовался набор алгоритмов Word2Vec в следующей последовательности: 1) определение частотности каждого слова в корпусе; 2) сортировка массива слов по частоте, удаление редких слов; 3) построение дерева Хаффмана (Huffman Binary Tree) для кодирования словаря (это значительно снижает вычислительную сложность алгоритма); 4) построение – с учетом заданного параметра окнá контекстов (максимальной дистанции между текущим и предсказываемым словом в предложении) – для каждого слова в корпусе вектора, элементы которого представляют собой обозначения количества случаев, когда данное слово оказывается в одном окне с другими наиболее частотными словами данного корпуса; 5) подача полученных векторов на вход нейросети прямого распространения (feedforward neural network), которая обучается предсказывать либо контекст по заданному слову, либо слово по заданному контексту. Векторное представление позволяет оценивать степень семантической близости каждой пары слов как косинусной меры их векторов, которая может принимать значения в промежутке [0, 1]: значение 1 – векторы слов ортогональны друг другу, у этих слов нет похожих контекстов и общих сем; значение 0 – практически полная идентичность контекстов, почти тождественная семантика слов. На основе различий векторных представлений лексем-терминов и лексем-лжетерминов был спроектирован алгоритм устранения некорректных кандидатов в термины (лжетерминов): карты взаимного расположения терминологических кандидатов в полученных дистрибутивно-

семантических моделях были спроецированы из векторного пространства высокой размерности в двухмерную плоскость, после чего была осуществлена кластеризация точек на плоскости по их координатам и маркировка каждого из полученных 20 кластеров как содержащего или не содержащего терминологическую лексику, произведенная с учетом ряда факторов (например, одним из факторов являлась удельная доля неоднословных сочетаний внутри кластера: предполагается, что в терминологических кластерах количество неоднословных единиц больше, поскольку автоматическое выделение терминологических сочетаний характеризуется более высоким уровнем точности, чем выделение отдельных терминов; подробнее см. [6]).

После осуществления описанной процедуры составленные списки извлеченных терминов по всем дисциплинам и всем уровням обучения были сопоставлены с действующими нормативными документами в области общего образования (федеральными образовательными стандартами и примерными программами основного общего и среднего общего образования); были сформированы также списки лжетерминов по каждому из подкорпусов. Для примера приведем некоторые статистические данные: (а) общее количество кандидатов в термины по всем подкорпусам (все дисциплины, все классы с 5-го по 11-й) – 48 911 единиц, из них: (б) терминов – 26282, (в) лжетерминов – 22 629, совпадений с терминологическими единицами в нормативных образовательных документах – 19 199. Аналогичные данные по некоторым учебным дисциплинам: алгебра – (а) 1 655, (б) 1526, (в) 129; биология – (а) 3 863, (б) 2 324, (в) 1539; всеобщая история и история России – (а) 5198, (б) 2491, (в) 2707; география – (а) 3352, (б) 1635, (в) 1717; геометрия – (а) 806, (б) 570, (в) 236; информатика – (а) 1294, (б) 682, (в) 612; русский язык – (а) 3780, (б) 2633, (в) 1147; физика – (а) 3254, (б) 2836, (в) 418; химия – (а) 2095, (б) 1087, (в) 288. Отношение терминов к числу всех лексем, употребляющихся в целевом корпусе с относительной частотой, значительно превышающей общеязыковую (т. е. терминов и лжетерминов), можно было бы условно обозначить как «коэффициент терминологической насыщенности». По этому коэффициенту учебники по разным учебным дисциплинам существенно отличаются друг от друга, при этом гуманитарные предметы отчетливо противопоставляются естественнонаучным и точным: для первых характерен низкий коэффициент (иначе говоря, значительная доля лжетерминов), для вторых – высокий коэффициент, т. е. решительное преобладание собственно терминологической лексики.

Все материалы, необходимые для воспроизведения результатов и верификации выводов статьи, кроме текстов учебников, являющихся собственностью правообладателя, размещены в постоянном научном хранилище Zenodo и доступны по адресу: https://zenodo.org/record/4079198#.X4Mrfy1h29Y.

Заключение. Проведенная работа показала прежде всего возможность применения алгоритмов Word2Vec и методов дистрибутивной се-

мантики для оптимизации результатов автоматического вычленения терминов из целевого корпуса, а именно для разделения массива автоматически выделенных лексем на собственно термины и слова, лишь уподобленные терминам в своем текстовом поведении. Отдельную проблему, которая может быть осмыслена в дидактическом отношении, составляет автоматически выявленное существенное число лжетерминов в учебниках по гуманитарным дисциплинам. В частности, монотонность содержания многих учебников по русскому языку, ощущаемая интуитивно, получает при реализованном подходе математически обоснованное подтверждение.

Дальнейшая реализация проекта предполагает сопоставление полученных данных с данными предобученных дистрибутивно-семантических моделей, предоставляемых сервисом RusVectōrēs (Национальный корпус русского языка и Википедия) [7], и данными модели, обученной на корпусе специальных научных текстов в тех сферах знания, которые представлены в школьных дисциплинах. Подобное сопоставление поможет оценить принципы организации терминов в векторном пространстве целевого корпуса и корпусов, содержащих аналогичные понятия в бытовой, научно-популярной и собственно научной сферах. Особыми направлениями в развитии проекта должны стать: а) автоматизация распределения терминов по тематических группам и соотнесение тематической структуры полученного перечня терминов с эксплицированной тематической структурой учебников; б) создание базы знаний по русской терминологической лексике, соответствующей содержанию общего образования в соответствии с федеральными стандартами; в) создание и обучение глубокой нейросети, способной по поданной на вход группе векторных представлений терминов определять учебную дисциплину, уровень обучения и учебную тему.

Помимо собственно теоретического значения в аспекте терминоведения, полученные и планируемые результаты могут иметь прикладную значимость при оценке эффективности школьного учебного текста, при составлении рекомендаций авторам школьных учебных пособий и при создании школьных терминологических словарей.

Список литературы

1. Kilgarriff A., Jakubíček M., Kovář V. et. al. Finding Terms in Corpora for Many Languages with the Sketch Engine // Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics, April 26-30, 2014. Gothenburg, 2014. Pp. 53–56.

2. Durda K., Buchanan L. WINDSORS: Windsor improved norms of distance and similarity of representations of semantics // Behavior Research Methods. 2008. Vol. 40. Pp. 705–712.

3. Jones M. N., Mewhort D. J. K. Representing word meaning and order information in a composite holographic lexicon // Psychological Review. 2007. Vol. 114. Pp. 1–37.

4. Mikolov T., Sutskever I., Chen K. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems 26. Cambridge, MA: MIT Press, 2013. Pp. 3111–3119.

5. Mikolov T., Yih W. T., Zweig G. Linguistic regularities in continuous space word representations // Human Language Technologies – North American Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2013. Pp. 746–751.

6. Монахов С. И., Турчаненко В. В., Федюкова Е. А., Чердаков Д. Н. Изучение терминологических подсистем современных школьных учебников на русском языке с помощью модели анализа семантики естественных языков Word2Vec // Journal of Applied Linguistics and Lexicography. 2020. Vol. 2. № 2. URL: https://journall.org/index.php/main/issue/view/4 (в печати).

7. Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science / Ignatov D. et al. (eds). 2017. Vol. 661. Pp. 155–161.