ГЕОЭКОЛОГИЯ

УДК 004.822:[502.52+639.2.053](261+265.1) doi: 10.33933/2713-3001-2021-63-279-296

Применение естественно-языкового машинного обучения для обобщения геоэкологических параметров пелагических биоресурсов Атлантического и Юго-Восточной части Тихого океанов

О.Ю. Краснобородько

Атлантический филиал ФГБНУ «ВНИРО» («АтлантНИРО»), Калининград,
sea@atlantniro.ru

Представлены результаты применения естественно-языковых технологий машинного обучения для анализа геоэкологической составляющей исследований промысловых биоресурсов в Атлантическом океане и в Юго-Восточной части Тихого океана. На основе текстов научных статей АтлантНИРО, содержащих результаты исследований за период 2007—2020 гг., в многомерном пространстве признаков смоделирована их понятийная область. В результате математических операций с координатными представлениями понятий обобщены и описаны геоэкологические факторы функционирования пелагических экосистем и их основных промысловых видов.

Ключевые слова: научные статьи, машинное обучение, обработка естественных языков, морская геоэкология, промысловые биоресурсы, геоэкологические факторы.

Application of natural language machine learning for generalization of geoecological parameters of pelagic biological resources in the Atlantic and South-Eastern Pacific oceans

O.Yu. Krasnoborodko

Atlantic branch of «VNIRO» («AtlantNIRO»), Kaliningrad, Russia

The article presents the results of the application of natural language processing in machine learning technologies, concerning the analysis of the geoecological component in studies of commercial biological resources in the Atlantic Ocean and southeastern Pacific Oceans. On the basis of lemmatized text corpus from 522 scientific articles containing the research results of the Atlantic branch of "VNIRO" for 2007–2020, its conceptual area is modeled in the multidimensional space of features within the principles of distributive semantics. Reducing the dimensions of space and visualizing the tags (sets of coordinates corresponding to the words of the corpus) were performed by assessing their compliance with the scientific researches (fishing, ichthyology, oceanology, hydroacoustics, hydrobiology, parasitology). The position of the word "Geoecology" in the tag-cloud has been determined and clusters of its associates have been identified — "Differentiation in time and space", "Unification", "Formation", "Transformation". As a result of mathematical operations with tags (calculation of combinations, proportions, correction by masks), the geoecological factors of the functioning of the pelagic ecosystems of the considered regions of the World Ocean and their main commercial species have been generalized and described. A number of nontrivial factors (spawning parameters, the influence of squids, space and geophysical factors, antiphase regime of commercial species, special

role of ecological optimum zones) as well as classical geoecological factors (dynamics of the atmosphere, currents, thermohaline characteristics of waters, the level of primary productivity) have been identified.

Keywords: scientific articles, machine learning, natural language processing, marine geoecology, commercial biological resources, geoecological factors.

For citation: *O.Yu. Krasnoborodko.* Application of natural language machine learning for generalization of geoecological parameters of pelagic biological resources in the Atlantic and South-Eastern Pacific oceans. *Gidrometeorologiya i Ekologiya.* Journal of Hydrometeorology and Ecology (Proceedings of the Russian State Hydrometeorological University). 2021. 63: 279—296. [In Russian]. doi: 10.33933/2713-3001-2021-63-279-296

Введение

Современные научно-исследовательские работы (НИР) — это деятельность с большим количеством степеней свободы. Они обусловлены как самим характером такой деятельности, так и ростом числа исследователей, современным широким доступом к научным данным и средствам их обработки, а также возможностями публикации и обмена результатами НИР. Особую сложность сегодня приобретают направления НИР, находящиеся на стыке наук, количество степеней свободы которых кратно количеству вовлеченных научных направлений. Одним из таких направлений является геоэкология — междисциплинарная наука, изучающая состав, структуру, закономерности функционирования и эволюции природных и антропогенных экосистем [1]. Современный лавинообразный процесс генерации и накопления данных, который получил название Big Data, уже сегодня не позволяет вести эффективный анализ вовлеченных в геоэкологию разнородных данных без применения средств их интеллектуальной обработки Data Mining, основанных, в первую очередь, на возможностях машинного обучения.

Машинное обучение — совокупность операций, в результате которых компьютерные системы постепенно улучшают качество выполнения задач без изменения базовых алгоритмов обработки данных. В середине 2000-х гг. бурное развитие получили технологии компьютерной обработки естественных языков Natural Language Processing (NLP), основанные на машинном обучении [2, 3]. В 2010-х гг. в связи с разработками Google [4], предложившей удобные программные инструменты трансформации текстов на естественных языках в наборы координат в многомерном пространстве признаков, поддающихся математическим операциям и компьютерной визуализации, NLP-исследования вышли на новый уровень. Это позволило компьютерным системам «понимать» естественный язык, а в Big Data машинного обучения вошел огромный массив неструктурированных данных — текстов из различных областей науки и техники, представленный главным образом виде научных статей.

Деятельность, связанная с исследованием состояния водных биологических ресурсов (ВБР), рассредоточенных в 66 крупных экосистемах Мирового океана [5], изначально носит геоэкологический характер [6]. В Атлантическом океане и в Юго-Восточной части Тихого океана (ЮВТО) ведущую роль в изучении океанических экосистем и ВБР районов исторического [7] и современного океанического промысла [8] играет Атлантический филиал ФГБНУ «ВНИРО» (далее — АтлантНИРО) (рис. 1).

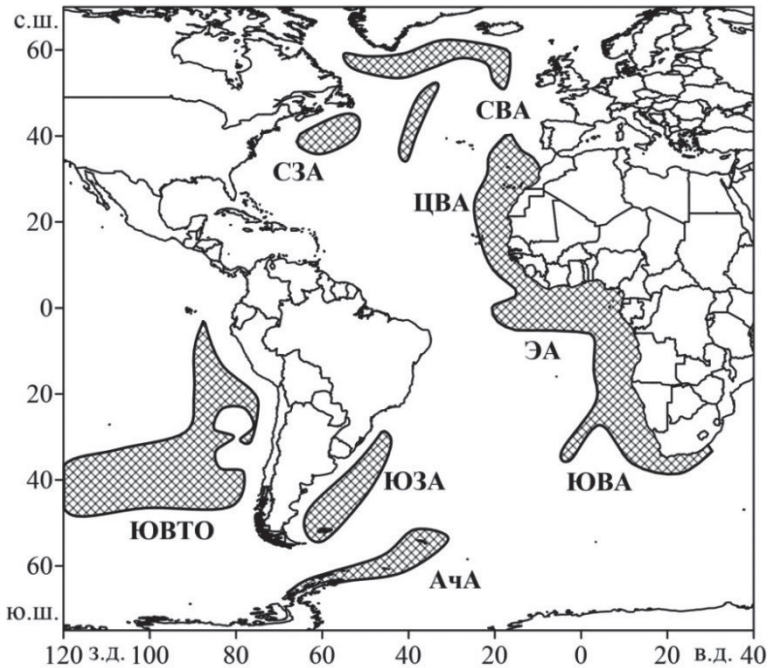

Рис. 1. Районы исторического и современного отечественного океанического промысла в Атлантике и ЮВТО, по [7, 8]:

СВА — Северо-Восточная Атлантика; СЗА — Северо-Западная Атлантика; ЦВА — Центрально-Восточная Атлантика; ЭА — Экваториальная Атлантика; ЮВА — Юго-Восточная Атлантика; ЮЗА — Юго-Западная Атлантика; АчА — Атлантическая часть Антарктики; ЮВТО — Юго-Восточная часть Тихого океана.

Fig. 1. Areas of historical and modern national oceanic fishing in the Atlantic and SEPO, according to [7, 8]:

NEA — North-Eastern Atlantic; NWA — North-Western Atlantic; CEA — Central-Eastern Atlantic; EA — Equatorial Atlantic; SEA — South-Eastern Atlantic; SWA — South-Western Atlantic; ApA — Atlantic part of Antarctica; SEPO —South-Eastern Pacific Ocean.

С середины 2000-х гг. большинство научных статей и обобщающих монографий [9, 10] по тематикам исследования ВБР и состояния условий среды в Атлантике и Тихом океане размещаются в Сети Интернет, в первую очередь в Российской научной электронной библиотеке ELibrary.Ru [11]. Учитывая имеющийся положительный опыт исследования естественнонаучных тематических корпусов (больших текстовых блоков) научных статей методами NLP как в России [2, 3, 12—15], так и за рубежом [16—18], массив статей АтлантНИРО также представляется перспективным полем для аналогичной исследовательской деятельности. В первую очередь, исследованию должен быть подвергнут наиболее сложный аспект выполняемых НИР, обусловленный самим характером предмета изучения — геоэкологическая составляющая исследований пелагических экосистем.

Целью данной работы является обобщение геоэкологических параметров функционирования пелагических экосистем в Атлантическом океане и ЮВТО, выявляемых в ходе исследований их ВБР на основе NLP-методов машинного обучения.

Материалы и методы

Текстовый корпус для обучения NLP-системы был сформирован из 522 научных статей на русском языке 130 авторов — сотрудников АтлантНИРО, опубликованных в 2007—2020 гг., которые размещены на портале ELibrary.Ru [11] в pdf-формате. Суммарный объем текстов всех статей составил 16,1 млн знаков, что эквивалентно 9 тыс. машинописных страниц. Иллюстрации, содержащиеся в статьях, не анализировались. На этапе предобработки тексты статей были объединены в единый текстовый корпус в txt-формате (рис. 2). С использованием инструмента MyStem, разработанного компанией Яндекс [19], текстовый корпус был подвергнут лемматизации, т. е. преобразованию всех его словоформ к их начальным формам (леммам), с удалением знаков препинания и преобразованием текста к нижнему регистру. На завершающем этапе предобработки из корпуса были удалены не имеющие смысловой нагрузки «стоп-слова» — предлоги, частицы, союзы.

Основная обработка обучающего текстового корпуса выполнялась авторским программным обеспечением, разработанным в среде Microsoft Visual Studio 2019 [20]. В качестве основного метода машинного обучения было выбрано обучение без учителя, а в основу алгоритма обработки данных был заложен основной принцип дистрибутивной семантики — лингвистические единицы в схожих контекстах имеют схожие значения [13]. Для всех лемм обучающего текстового корпуса на основе метода «мешок слов» [4] с контекстным окном шириной 50 слов автоматически формировались узлы единого орграфа, которые в процессе обучения системы видоизменялись. В конце обучения общей длительностью 240 машино-часов орграф содержал 26,7 тыс. узлов, связанных почти 4 млн дуг-связей. Каждый узел орграфа имел до 8 тыс. дуг (входов и выходов) с весами от 0 до 1, отражающими степень близости леммы (стоящего за ней слова) с другими словами в контекстах обработанных научных статей.

На основе орграфа для каждого слова рассчитывался набор его координат в многомерном пространстве признаков — тег (авт.). В отличие от векторного понимания слов текста, подразумевающего пары координат для каждого слова и векторные математические операции с ними [4], тег имеет только один набор координат, т. е. представляет собой скаляр в многомерном пространстве и менее ограничен в выборе методов его обработки. Количество измерений в этом пространстве равно количеству узлов в орграфе, т. е. объему словаря текстового корпуса.

Для понижения размерности признакового пространства с целью его визуализации и дальнейшего анализа в трехмерном пространстве применялся метод взвешивания тегов по степени их соответствия основным тематикам НИР

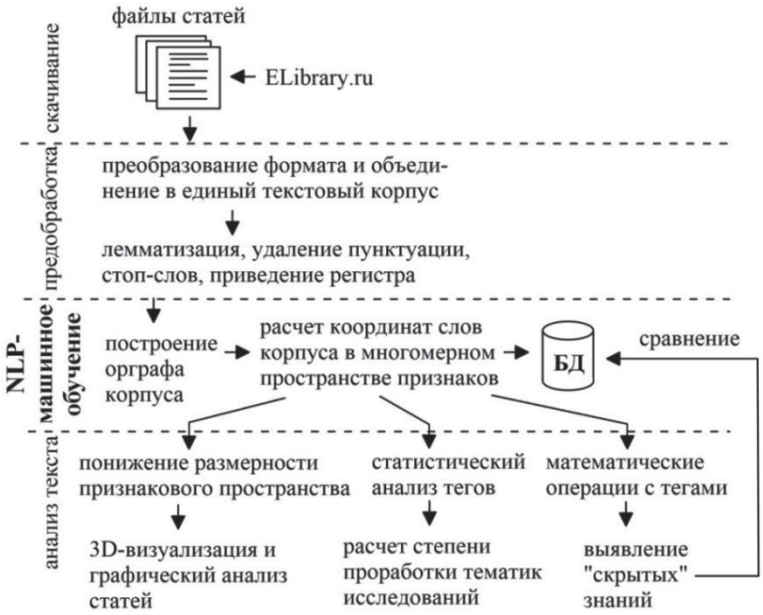

Рис. 2. Этапы обработки и анализа данных 522 научных статей на русском языке.

Fig. 2. Stages of data processing and analysis of 522 scientific articles in Russian.

АтлантНИРО. Ключевые понятия тематик (промысел, ихтиология, океанология, гидроакустика, гидробиология, паразитология) были заранее распределены в пространстве «X, Y, Z» равноудалено от начала координат. Для нивелирования размерности получаемого трехмерного пространства по осям применялось его вращение вокруг оси Z. Визуализация и графический анализ текстов статей, преобразованных таким образом в наборы трехмерных тегов, выполнялись в программах Surfer и Voxler [21].

Полученные координатные представления слов позволили провести статистический анализ текстов научных статей для оценки геоэкологической составляющей исследований океанических биоресурсов, проводимых в Атлантике и ЮВТО. Также они дали возможность выполнить математические операции с различными комбинациями научных понятий, входящими в статьи. Возможность таких операций была показана ранее [4] и является эффективным инструментом для выявления «скрытых» взаимосвязей понятий, содержащихся в анализируемых методами NLP текстах.

Для выявления «скрытых» факторов функционирования пелагических экосистем исследуемых районов применялись два типа координатных операций с тегами в многомерном пространстве признаков.

1 операция — расчет средних координат комбинаций тегов, их коррекция по маскам геоэкологических тегов (рис. 3):

$$расчетный_тег = \frac{район_i + "фактор" + "влияние"}{3}, если_(гео - тег_j \neq 0), \quad (1)$$

где *расчетный_тег* — координаты тега искомого геоэкологического фактора для района *i* (1 — «Роколл», 2 — «Ньюфаундленд», 3 — «Канарский», 4 — «Гвинейский», 5 — «Бенгельский», 6 — «Фолклендский», 7 — «Антарктический», 8 — «Перуанский»); $район_i$ — координаты тега-названия района *i*; «фактор», «влияние» — координаты тегов-констант; $гео\text{-}тег_j$ — координаты геоэкологического тега-маски *j*, по которому корректируется расчетный тег (1 — «экология», 2 — «геоэкологический», 3 — «экосистема»).

2 операция — расчет пропорций координат тегов:

$$\frac{расчетный_тег}{"влияние"} = \frac{фактор_i}{вид_j}, \quad (2)$$

где *расчетный_тег* — координаты тега искомого геоэкологического фактора для промыслового вида *j* (1 — «анчоус», 2 — «криль, креветка», 3 — «сардина», 4 — «сардинелла», 5 — «скумбрия», 6 — «ставрида», 7 — «окунь», 8 — «сардинопс», 9 — «кальмар», 10 — «путассу», 11 — «тунец»); $фактор_i$ — координаты тега геоэкологического фактора *i* (1 — «шельфовый», 2 — «звукорассеивающий», 3 — «адвекция», 4 — «физиологический», 5 — «метеорологический» , 6 — «термохалинный», 7 — «перенос»); «влияние» — координаты тега-константы; $вид_j$ — координаты тега-названия промыслового вида *j*.

Вес выявленных геоэкологических факторов по районам исследований (%) рассчитывался как величина, обратная расстоянию между расчетными тегами и наиболее близко расположенными к ним тегами облака, несущими

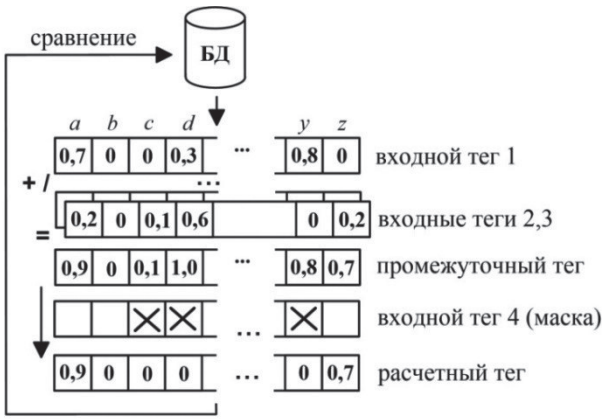

Рис. 3. Пример координатных операций с тегами:

a—z — координаты тега в многомерном пространстве признаков; БД — база данных.

Fig. 3. Example of coordinate operations with tags:

a—z — coordinates of the tag in the multidimensional space of features; DB — database.

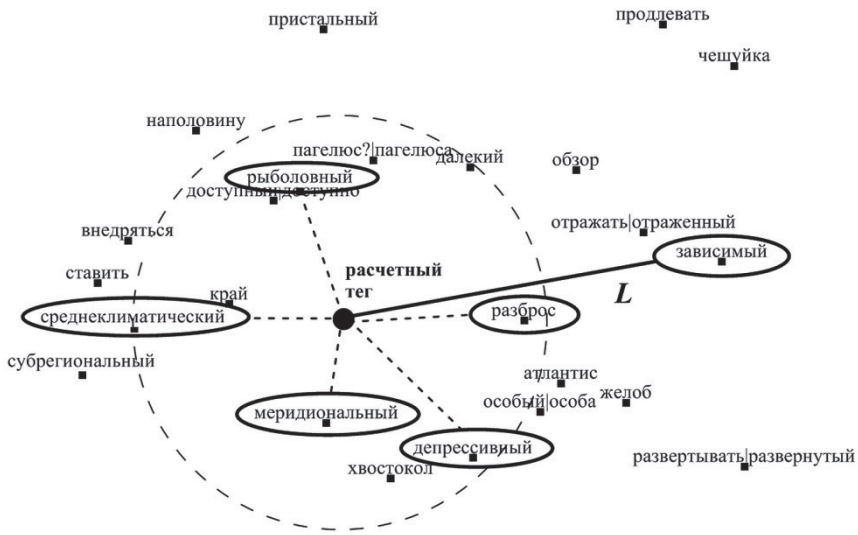

Рис. 4. Пример выделения геоэкологических тегов в области расчетного тега, оценка расстояния до тега «Зависимый» (L).

Fig. 4. An example of mining of geoecological tags in the area of the calculated tag, estimating the distance to the tag “Dependent” (L).

геоэкологический смысл. Приоритетность геоэкологических факторов по видам гидробионтов оценивалась по наименьшим расстояниям между расчетными тегами и тегом «зависимый» (рис. 4).

Результаты

В результате обработки текстового корпуса было получено распределение его тегов (понятий) в трехмерном пространстве тематик НИР АтлантНИРО, определено положение в нем термина «Геоэкология», оказавшегося на стыке всех рассматриваемых направлений НИР с некоторым уклоном в сторону направлений НИР «Промысел», «Ихтиология» и «Океанология» (рис. 5).

В области нахождения термина «Геоэкология» были выявлены несколько групп (кластеров) его ассоциатов — сочетаний терминов, возникающих в оперативной памяти компьютера в ответ на поданный на вход NLP-системы другой термин. Ассоциаты объединены с ним близкой смысловой нагрузкой (рис. 6):

— кластер 1 — группа понятий «Дифференциация во времени и пространстве»;

— кластер 2 — «Объединение»;

— кластер 3 — «Формирование»;

— кластер 4 — «Трансформация».

В результате выполнения операции (1) с комбинациями тегов были получены новые расчетные теги, положение которых в трехмерном пространстве тематик

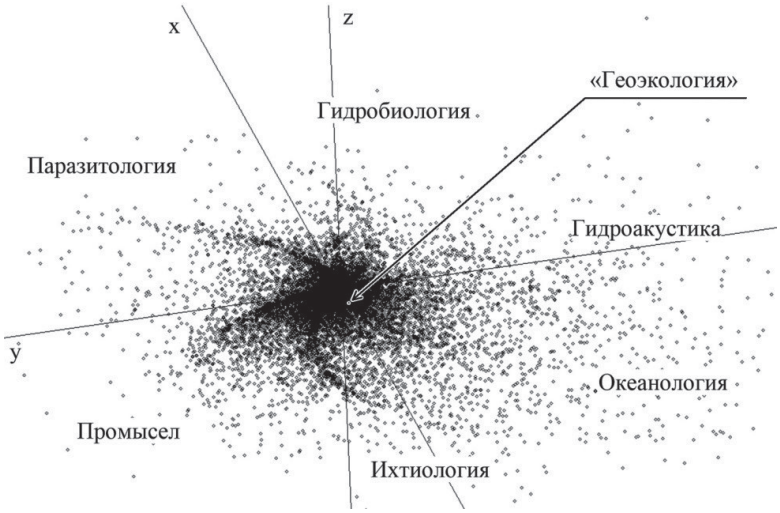

Рис. 5. 3D-облако тегов из статей АтлантНИРО
в пространстве тематик НИР и положение тега «Геоэкология».

Fig. 5. 3D-tag cloud from AtlantNIRO articles in the space of research topics and position of the tag «Geoecology».

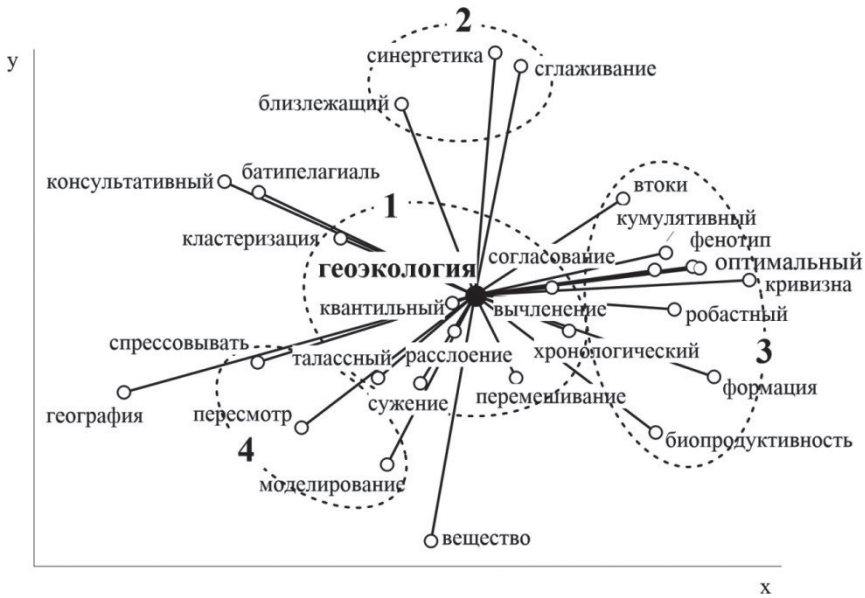

Рис. 6. Ассоциаты термина «Геоэкология» в проекции облака на плоскость «XY»:
1—4 — кластеры ассоциатов.

Fig. 6. Associates of the term “Geoecology” in the projection of the cloud on the plane “XY”:
1—4 — clusters of associates.

НИР было близко к терминам, характеризующими экологические факторы развития океанических экосистем в исследуемых районах (табл. 1). Вес выявленных геоэкологических факторов варьирует в широких пределах (от 100 до 10 %) и отражает степень изученности фактора в рамках геоэкологической науки.

Выполнение операции (2) с комбинациями тегов «промысловый вид» — «геоэкологический фактор» позволило выделить приоритетные геоэкологические факторы влияния на основные массовые промысловые виды гидробионов Атлантического океана и ЮВТО (табл. 2). Количество таких факторов у разных видов варьирует от одного (окунь) до трех (сардина). Отмечается соответствие выделенных факторов экологической роли видов (фитофаги, зоофаги и хищники).

Таблица 1

Геоэкологические факторы функционирования пелагических экосистем Атлантики и ЮВТО, выявленные операцией (1) (в скобках курсивом их ассоциаты)

Geoecological factors of the functioning of the pelagic ecosystems of the Atlantic Ocean and SEPO, identified by operation (1) (in parentheses their semantic associates)

Район, подрайон	Фактор (*смысловой ассоциат*)	Вес, %
СВА, возвышенность Роколл	адвективный (*поток*)	100
	звукорассеивающий (*слой*)	83
	локальный (*район, процесс*)	75
	дневной (*период*)	39
	весенний (*сезон*)	33
	физиологическое (*состояние*)	25
	вертикальный (*профиль*)	17
	термический (*фон*)	17
СЗА, Ньюфаундленд	звукорассеивающий (*слой*)	83
	метеорологический (*режим*)	63
	поверхностный (*слой*)	57
	адвективный (*поток*)	31
	осенне-зимний (*период*)	19
	вертикальный (*профиль*)	17
	термохалинные (*характеристики*)	14
ЦВА, Канарская экосистема	смешанный (*скопление*)	100
	депрессия (*барическая*)	99
	устойчивый (*апвеллинг, пассат*)	93
	приливной (*фронт, волны*)	67
	стайный (*вид*)	50
	интрузионный (*поток*)	50
	миграционный (*путь, период*)	30
ЭА, Гвинейский залив	адвективный (*поток*)	92
	звукорассеивающий (*слой*)	85
	вертикальный (*профиль*)	75
	прибрежный (*район*)	67
	смежный (*район*)	54
	разрывной (*поток, течение*)	54
	внутренний (*структура*)	50
	физиологический (*состояние*)	28

Окончание табл. 1

Район, подрайон	Фактор (*смысловой ассоциат*)	Вес, %
ЮВА, Бенгельская экосистема	термический (*режим*)	100
	вертикальный (*профиль*)	98
	метеорологический (*режим*)	59
	«окно» (*оптимума, заселения*)	56
	адвективный (*поток*)	31
	весенний (*период*)	29
ЮЗА, Фолклендский подрайон	подвижность (*течения, фронта, скопления*)	100
	скумбрия (*вид*)	40
	шельфовый (*район*)	40
	скопление (*гидробионтов*)	22
	противофазный (*режим*)	20
АчА, Антарктический полуостров	смещение (*скоплений, фронта*)	67
	климатический (*режим*)	64
	циркуляционный (*режим, картина*)	52
	тропосферный (*слой*)	50
	стратификация (*вертикальная*)	44
	космогеофизический (*фактор*)	25
	термохалинный (*режим*)	16
ЮВТО, Перуано-Чилийский подрайон	течение (*океаническое*)	58
	западный (*перенос, течение*)	58
	сеголетки (*молодь*)	56
	оммастрефиды (*кальмары*)	54
	космогеофизический (*фактор*)	38
	ветровой (*режим*)	22
	нерестовый (*период, состояние*)	10

Таблица 2

Приоритетные геоэкологические факторы для основных видов промысловых гидробионтов Атлантики и ЮВТО, выявленные операцией (2)

Priority geoecological factors for the main species of commercial aquatic organisms in the Atlantic Ocean and SEPO, identified by operation (2)

	Батиметрия	Состояние ЗРС	Динамика вод	Физиологическое состояние	Метеорологические условия	Термохалинные характеристики	Миграции вида
Анчоус	+	+	—	—	—	—	—
Криль	+	+	—	—	—	—	—
Сардина	+	+	+	—	—	—	—
Сардинелла	+	+	—	—	—	—	—
Скумбрия	—	—	+	+	—	—	—
Ставрида	—	+	—	—	+	—	—
Окунь	—	—	+	—	—	—	—
Сардинопс	—	—	—	+	—	+	—
Кальмар	—	—	—	+	—	+	—
Путассу	—	—	—	+	—	—	+
Тунец	—	—	—	—	+	—	+

Обсуждение

Полученное в результате NLP-обработки текстов статей распределение их тегов в многомерном пространстве признаков позволило по-новому взглянуть на проблему поиска новых геоэкологических знаний на базе ранее опубликованных результатов НИР АтлантНИРО. Охватываемые в статьях тематики НИР и их наполненность научными фактами, терминологией, внутренними и внешними смысловыми связями позволяют проводить их эффективный поиск существующими в рамках NLP-подхода методами.

Одним из ключевых вопросов технологии NLP является визуализация раскрываемых многомерных пространств текстов. От качества визуализации и принципов, которые в нее заложены, напрямую зависят возможности последующего анализа этой информации. Сегодня существует большое количество способов визуализации многомерных пространств. В примененном нами подходе изначально заложен базовый принцип дифференциации направлений НИР (см. рис. 5). В плотном центре (ядре) облака тегов концентрируются малозначимые с научной точки зрения общеупотребимые слова («исправлять», «сочетаться» и др.). В окружающем его более разреженном пространстве в основном локализуются научные термины междисциплинарного характера («отчетность», «логарифм», «подекадный» и др.). В спиралеподобных «рукавах» скопления расположились главным образом специфические термины, употребляемые преимущественно в рамках конкретных направлений НИР — причем, чем дальше термин находится от общего для всех направлений ядра, тем он более специфичен («самец», «нереститься», «популяционный» и др.).

Тег термина «Геоэкология» ожидаемо попал в центральную область облака, где в основном группируются термины, лежащие на стыке научных направлений. Будучи пограничным научным направлением, геоэкология вбирает в себя методики и знания разных направлений НИР. Некоторый сдвиг тега отмечен в сторону т. н. биологических направления исследований — «Океанология», «Ихтиология», «Промысел», у которых отмечена высокая степень проработки и тесная интеграция результатов исследований, что выражается в большом количестве тегов между «рукавами» этих направлений (см. рис. 5).

Тематические корпуса научных статей содержат множественные явные и «скрытые» информационные связи, как ближние тематические (со смысловыми единицами в рамках конкретного направления исследований), так и дальние смысловые (выходящие за рамки отдельных направлений НИР, междисциплинарные). Степень связанности понятий определяется их близостью в контекстах.

Примененный NLP-подход обнаружил естественную (заложенную в смысл понятий) группировку ассоциатов термина «Геоэкология» на четыре кластера (см. рис. 6). Особым среди них можно считать кластер 1, содержащий сам термин «Геоэкология». Термины, находящиеся с ним в этом же классе, несут близкую смысловую нагрузку, которая, вероятно, характеризует и саму «Геоэкологию» как науку. Они дифференцируют явления во времени (ассоциат «хронологический»), пространстве (ассоциаты «расслоение», «вычленение») и даже в многомерных

математических пространствах («кластеризация», «квантильный»). При этом дифференциация не является конечным результатом, а имеет и обратный механизм объединения и сглаживания (кластер 2), трансформации (кластер 4), формирования и закрепления нового (кластер 3). Наряду с общенаучными терминами в геоэкологические кластеры попали биологические и океанологические термины («фенотип», «талассный», «втоки», «биопродуктивность»), что указывает на тесную связь между этими направлениями исследований (см. рис. 6).

Выявление «скрытых» связей и новых знаний, не содержащихся в явном виде в научных работах, представляет особый интерес. Такие данные могут быть выявлены в ходе математических операций с полученными из статей тегами, в сжатой цифровой форме несущими их контексты. Следует отметить, что результаты математических операций с тегами всегда носят вероятностный характер, теги быстро «деградируют» (теряют смысл) с каждым последующим шагом обработки, а «ответы» системы всегда значительно зашумлены в силу следующих причин.

1. Недостаточный объем обучающей выборки. В NLP-машинном обучении обучающие выборки обычно строятся из миллионов текстов. Однако при таком подходе, на наш взгляд, в большом текстовом потоке могут нивелироваться немассовые, уникальные, штучные элементы знаний, содержащихся в небольших фрагментах текстов. В этом смысле, апробация NLP-методов на малом тематическом корпусе статей АтлантНИРО представляет интерес.

2. Изначально низкое качество некоторой части научных статей, содержащих либо «слабые», либо непроверенные, либо ошибочные научные результаты.

3. Смена идей, изложенных в статьях. Несмотря на то, что смена идей — это естественный процесс, машинное обучение на таких корпусах, вероятно, приведет к усвоению NLP-системой и новых и старых (ошибочных) знаний.

4. Сложность самих языковых явлений. Естественный язык отличается сложной, контекстно-зависимой семантикой, в научных статьях ее уровень высок, и это создает дополнительные сложности при построении NLP-систем.

Анализ выявленных наборов геоэкологических факторов функционирования пелагических экосистем промысловых районов Атлантического океана и ЮВТО показывает, что для каждого района он уникален (см. табл. 1). При этом, наряду с такими известными факторами, как динамика атмосферы, течения, термохалинные характеристики вод, состояние первичной продуктивности района (параметры т. н. звукорассеивающих слоев — ЗРС) и другими, выявился ряд нетривиальных факторов, первичная проверка некоторых показала их фактическое существование:

1) состояние ВБР в нерестовый период, состояние сеголеток и влияние на пелагическую экосистему кальмаров в ЮВТО;

2) влияние космогеофизических факторов на экосистему АчА;

3) противофазный режим состояния ВБР кальмаров и скумбрии в ЮЗА;

4) особая роль «окон» (вероятно, «заселения») в пелагической экосистеме ЮВА.

В системе координат геоэкологических факторов основные промысловые виды правильно ранжируются от типичных фитофагов (анчоус, криль, сардина) до зоофагов и хищников (кальмар, путассу, тунец). Для первых приоритетными

факторами являются батиметрия района обитания (наличие шельфа для рыб и его отсутствие для криля) и стояние ЗРС (кормовой базы). Для вторых приоритетными являются термохалинные характеристики вод и благоприятные условия для миграций (см. табл. 2).

Координатные представления больших текстовых корпусов в едином пространстве признаков позволяют охватить их одним взглядом, одномоментно. В рамках геоэкологического подхода к НИР в Атлантике и ЮВТО это позволяет обобщить ведущие геоэкологические факторы функционирования пелагических экосистем всех рассмотренных районов (рис. 7).

В СВА (подводная возвышенность Роколл и примыкающая к ней с севера зона Фарерских островов) основными промысловыми видами являются путассу и скумбрия. В функционировании пелагической экосистемы района, наряду с другими традиционными факторами, существенную роль играют адвекция океанических субполярных вод на шельф Ирландии и Великобритании [7]. Проявление этого процесса в каждом из подрайонов носит локальный характер и определяет условия общей фоновой биопродуктивности вод в нем.

В СЗА (подрайон Ньюфаундленда, Новой Шотландии, подводной возвышенности Флемиш-Кап) основными промысловыми видами являются окунь, креветка, кальмар). Район является динамически активным вследствие взаимодействия холодного Лабрадорского течения и теплого течения Гольфстрим [7]. В СЗА значительно влияние метеорологического режима (это район активной циклонической деятельности), большое влияние на пелагическую экосистему района оказывают поверхностные процессы взаимодействия холодных северных и теплых южных вод, а также процессы, регулирующие общий уровень биологической продуктивности и состояние ЗРС.

Скумбрия, ставрида, сардина и сардинелла составляют основу промыслового ихтиоцена ЦВА в пределах экономических зон Марокко, Мавритании, Сенегала, Республики Гвинеи-Бисау. К важнейшим факторам функционирования Большой Канарской пелагической экосистемы относятся поддержание здесь зоны пассатов, разделяющих крупные барические образования региона — Азорский антициклон и Сахарскую депрессию, формирование прибрежного апвеллинга, а также смешанный характер скоплений массовых видов рыб, максимально плотно заселяющих существующие экологические ниши [7—9].

Для тунцов Экваториальной Атлантики (Гвинейский залив, открытая часть Атлантики) в их сезонном миграционном цикле крайне важными являются параметры вергентных зон, меняющие стратификацию вод (степень заглубления термоклина — важного средообразующего параметра для разных видов тунцов). Не менее важным для района оказывается общий уровень биопродуктивности вод, который можно оценить по состоянию ЗРС [7].

Для района ЮВА (экономические зоны Анголы, Намибии) важными регуляторами состояния пелагиали являются те же факторы, которые работают в ЦВА — пассат, поддерживающий прибрежный апвеллинг, параметры «окон» заселения видов (экологические ниши в водных массах). Основными добываемыми видами здесь являются ставрида, скумбрия, сардинопс [7, 8].

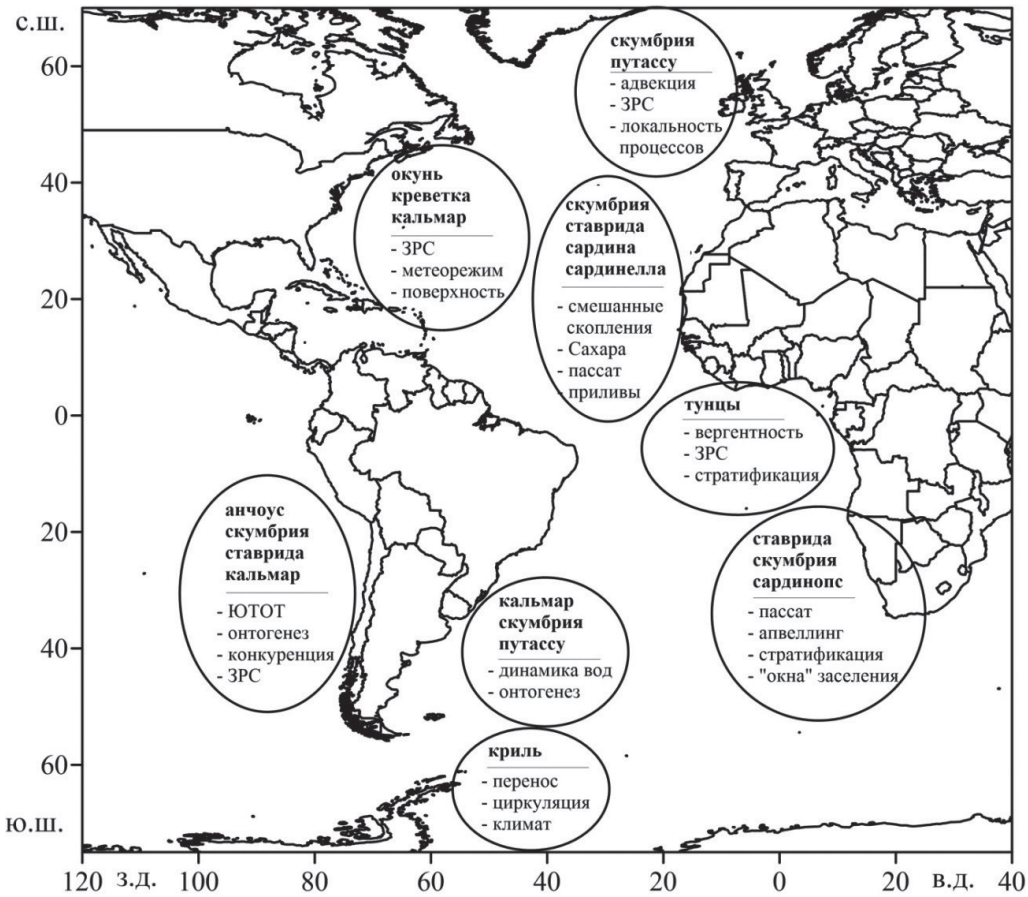

Рис. 7. Ведущие геоэкологические факторы в районах НИР, выявленные по результатам NLP-обработки текстового корпуса статей АтлантНИРО 2007—2020 гг. (пояснения в тексте).

Fig. 7. Leading geoecological factors in research areas, identified on the base of NLP with text corpus of AtlantNIRO articles 2007—2020 (see the text explanations).

В ЮЗА (Аргентинский, Фолклендский подрайоны) промысел в основном базируется на запасах кальмаров (аргентинский, лолиго), так же облавливаются пелагические виды рыб (южная путассу, скумбрия). Район отличается сложной динамикой вод во взаимодействующих холодном Фолклендском и теплом Бразильском течениях. Также для пелагической экосистемы района характерна многолетняя цикличность обилия кальмаров и рыб, вызывающая вспышки и падения их численности [7, 8].

Пелагическая экосистема АчА — основной район промысла антарктического криля — привлекает большое внимание как поле будущей научной и промысловой деятельности России и других стран в Антарктике. Основные факторы ее

функционирования хорошо известны — перенос скоплений криля в пределах Антарктического циркумполярного течения, параметры атмосферной и океанической циркуляции, долгопериодные климатические колебания [7, 8].

В ЮВТО, где добываются основные мировые запасы анчоуса, в прошлом велся активный промысел ставриды и скумбрии, а сегодня — кальмара, пелагическая экосистема регулируется параметрами течений открытого океана, прежде всего, Южно-Тихоокеанского океанического течения (ЮТОТ), которое определяет уровень биопродуктивности открытых вод района [7, 8, 10]. В ходе воспроизводства популяции основного вида района — ставриды — немаловажную роль играют факторы успешности нереста, раннего онтогенеза пополнения и пищевой конкуренции с головоногими.

Заключение

Морские геоэкологические исследования — это деятельность с большим количеством степеней свободы. Успешному управлению и анализу результатов этой деятельности может способствовать видение процессов и результатов НИР в целом. Такую возможность дают современные методы интеллектуальной обработки и анализа текстов, основанные на NLP.

Полученное в результате NLP-обработки текстов статей АтлантНИРО 2007—2020 гг. распределение их тегов в многомерном пространстве признаков позволило по-новому взглянуть на проблему поиска и анализа содержащихся в них геоэкологических знаний. В основе примененных в работе новых информационных технологий лежит работа с семантической (смысловой) нагрузкой содержащихся в них терминов. Полученные в ходе исследования результаты не противоречат существующим взглядам на факторы функционирования пелагических экосистем Мирового океана, однако детализируют их под другим углом зрения.

В ходе исследования выявлены ассоциативные связи геоэкологии с другими ключевыми понятиями из области океанологии, математической статистики, биологии и др. Получено адекватное автоматическое ранжирование основных промысловых видов всех рассмотренных подрайонов по их экологической роли — от фитофагов до зоофагов и хищников. Анализ наборов геоэкологических факторов функционирования пелагических экосистем по районам промысла в Атлантическом океане и ЮВТО, а также по промысловым видам показал, что для каждого района и промыслового вида он уникален. В рамках геоэкологического подхода к НИР АтлантНИРО в изучаемых районах обобщен набор факторов и рассмотрены механизмы их влияния на пелагические экосистемы.

В целом, NLP-методы машинного обучения являются перспективным средством анализа геоэкологической составляющей современных НИР, выполняемых в промысловых районах Атлантики и ЮВТО. Научный потенциал АтлантНИРО в системе векторов развития, задаваемых требованиями к проведению геоэкологических ресурсных морских изысканий, по всем районам исследований является высоким. Об этом свидетельствует высокая степень проработки и интеграции тематик НИР АтлантНИРО, особенно в рамках биологического направления исследований — основного поля деятельности морской геоэкологии.

Целесообразно продолжать работы по построению NLP-систем, выявлению, обобщению и последующей интерпретации профильными специалистами т. н. «скрытых» знаний, содержащихся в массивах ранее опубликованных научных работ.

Полученные в работе результаты также могут быть полезны при планировании последующих НИР по исследованию ВБР в промысловых районах Атлантического океана и ЮВТО, как ориентиры наиболее функционально нагруженных компонентов входящих в них пелагических экосистем.

Список литературы

1. *Трофимов В.Т., Зилинг Д.Г., Аверкина Т.А.* Геоэкология как термин и междисциплинарная наука // Вестник МГУ. Сер. 4. Геология. 1994. № 5. С. 43—55.
2. *Белов С.Д., Зрелова Д.П., Зрелов П.В., Кореньков В.В.* Обзор методов автоматической обработки текстов на естественном языке // Системный анализ в науке и образовании: сетевое научное издание. 2020. № 3. С. 8—22.
3. *Найденова К.А., Невзорова О.А.* Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований // Ученые записки Казанского государственного университета. Т. 150. Кн. 4. Физико-математические науки. 2008. С. 5—24.
4. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // In Proceedings of Workshop at ICLR. 2013. P. 1—12.
5. *Belkin I.* Rapid warming of Large Marine Ecosystems // Progress in Oceanography. 2009. 81 (1-4). P. 207—213. doi: 10.1016/j.pocean.2009.04.011.
6. *Холмянский М.А., Снопова Е.М., Владимиров М.В., Заносов О.Ю.* Современное состояние и прогноз развития морских геоэкологических исследований // Ученые записки РГГМУ. Вып. 48. 2017. С. 216—233.
7. Промыслово-океанологические исследования в Атлантическом океане и южной части Тихого океана (по результатам исследований АтлантНИРО и Запрыбпромразведки) / Под. ред. В.Н. Яковлева. Калининград: Изд. АтлантНИРО, 2002. Т. 1, 2. 521 с.
8. *Бандурин К.В., Архипов А.Г.* Современное состояние водных биологических ресурсов в океанических районах исследований АтлантНИРО // Ученые записки РГГМУ. Вып. 48. 2017. С. 148—166.
9. *Малинин В.Н., Чернышков П.П., Гордеева С.М.* Канарский апвеллинг: крупномасштабная изменчивость и прогноз температуры воды. СПб.: Гидрометеоиздат, 2002. 155 с.
10. *Малинин В.Н., Гордеева С.М.* Промысловая океанология юго-восточной части Тихого океана. Том. I. Изменчивость факторов среды обитания. СПб.: РГГМУ, 2009. 278 с.
11. Российская электронная научная библиотека. [Электронный ресурс]. URL: https:elibrary.ru (дата обращения: 27.02.2021).
12. *Башков А.С., Соломенцев Я.К.* Использование векторных методов представления слов в задачах выявления трендов // Вестник Российского нового университета. 2019. № 2. С. 80—88. doi: 10.25586/rnu.v9187.19.02.p.080.
13. *Трофимов И.В., Сулейманова Е.А.* Дистрибутивно-семантическая модель для выявления категориального сходства // Программные системы: теория и приложения. Т. 9. № 4 (39). 2018. С. 443—460. doi: 10.25209/2079-3316-2018-9-4-443-460.
14. *Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.* Автоматическая обработка текстов на естественном языке и анализ данных. М.: НИУ ВШЭ, 2017. 269 с.
15. *Лукашевич Н.В., Добров Б.В.* Проектирование лингвистических онтологий для информационных систем в широких предметных областях // Онтология проектирования. 2015. Т. 5, № 1 (15). С. 47—69.
16. *Lima E.B., Davis C.A.* Geographic information extraction using Natural Language Processing in Wikipedia texts // Proceedings XVIII GEOINFO. 2017. P. 122—127.

17. *Qiu Q., Xie Z., Wu L., Tao L.* Dictionary-based automated information extraction from geological documents using a Deep Learning Algorithm // Earth and Space Science. 7. 2020. P. 1—18. doi: 10.1029/2019EA000993.
18. *Wang C., Ma X., Chen J., Chen J.* Information extraction and knowledge graph construction from geoscience literature // Computers and Geosciences. V. 112. 2018. P. 112—120. doi: 10.1016/j.cageo.2017.12.007.
19. MyStem. [Электронный ресурс]. URL: https://yandex.ru/dev/mystem (дата обращения: 28.02.2021).
20. Microsoft Visual Studio 2019. [Электронный ресурс]. URL: https://visualstudio.microsoft.com/ru (дата обращения: 28.02.2021).
21. Surfer. Voxler. Golden Software Inc. [Электронный ресурс]. URL: https://www.goldensoftware.com/products (дата обращения: 28.02.2021).

References

1. *Trofimov V.T., Ziling D.G., Averkina T.A.* Geoecology as a term and interdisciplinary science. *Vestnik MGU. Ser. 4. Geologiia*. Bulletin of Moscow State University. Part 4. Geology. 1994, 5: 43—55. [In Russian].
2. *Belov S.D., Zrelova D.P., Zrelov P.V., Korenkov V.V.* Review of methods of automatic processing of texts in natural language. *Sistemnyi analiz v nauke i obrazovanii: setevoe nauchnoe izdanie*. System analysis in science and education: a network scientific publication. 2020, 3: 8—22. [In Russian].
3. *Naydenova K.A., Nevzorova O.A.* Machine learning in natural language processing problems: an overview of the current state of research. *Uchenye zapiski Kazanskogo gosudarstvennogo universiteta. T. 150. Kn. 4. Fiziko-matematicheskie nauki.* Scientific notes Kazan State University. T. 150. Book. 4. Physical and mathematical sciences. 2008: 5—24. [In Russian].
4. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // In Proceedings of Workshop at ICLR. 2013: 1—12.
5. *Belkin I.* Rapid warming of Large Marine Ecosystems. Progress in Oceanography. 2009, 81 (1-4): 207—213. doi: 10.1016/j.pocean.2009.04.011.
6. *Kholmyansky M.A., Snopova E.M., Vladimirov M.V., Zanosov O.Yu.* Current state and forecast of development of marine geoecological research. *Uchenye zapiski RGGMU.* Scientific notes RSHU. 2017, 48: 216—233. [In Russian].
7. *Promyslovo-okeanologicheskie issledovaniia v Atlanticheskom okeane i iuzhnoi chasti Tikhogo okeana (po rezul'tatam issledovanii AtlantNIRO i Zaprybpromrazvedki).* Commercial oceanological research in the Atlantic Ocean and the South Pacific Ocean (based on the results of AtlantNIRO and Zaprybpromrazvedka research). Edited by V.N. Yakovlev. Kaliningrad: AtlantNIRO, 2002. T. 1, 2: 521. [In Russian].
8. *Bandurin K.V., Arkhipov A.G.* The current state of aquatic biological resources in the oceanic research areas of AtlantNIRO. *Uchenye zapiski RGGMU.* Scientific notes RSHU. 2017, 48: 148—166. [In Russian].
9. *Malinin V.N., Chernyshkov P.P., Gordeeva S.M. Kanarskii apvelling: krupnomasshtabnaia izmenchivost' i prognoz temperatury vody.* Canary upwelling: large-scale variability and forecast of water temperature. SPb.: Gidrometeoizdat, 2002: 155. [In Russian].
10. *Malinin V.N., Gordeeva S.M. Promyslovaia okeanologiia iugo-vostochnoi chasti Tikhogo okeana. Tom. I. Izmenchivost' faktorov sredy obitaniia*. Fishery oceanology of the southeastern part of the Pacific Ocean. Tom I. Variability of environmental factors. SPb.: RSHU, 2009: 278. [In Russian].
11. Russian electronic scientific library. Available at: https:elibrary.ru (date of access: 27.02.2021).
12. *Bashkov A.S., Solomentsev J.K.* Using vector methods for representing words in the tasks of identifying trends. *Vestnik Rossiiskogo novogo universiteta*. Bulletin of the Russian New University. 2019, 2: 80—88. doi: 10.25586 / rnu.v9187.19.02. 080. [In Russian].
13. *Trofimov I.V., Suleimanova E.A.* Distributive-semantic model for identifying categorical similarities. *Programmnye sistemy: teoriia i prilozheniia. T. 9.* Software systems: theory and applications. T. 9. 2018, 4(39): 443—460. doi: 10.25209/2079-3316-2018-9-4-443—460. [In Russian].
14. *Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaia obrabotka tekstov na estestvennom iazyke i analiz dannykh*. Automatic natural language processing and data analysis. M.: NRU HES, 2017: 269. [In Russian].

15. *Lukashevich N.V., Dobrov B.V.* Designing linguistic ontologies for information systems in broad subject areas. *Ontologiia proektirovaniia*. Design Ontology. 2015. Vol. 5, 1 (15): 47—69. [In Russian].
16. *Lima E.B., Davis C.A.* Geographic information extraction using Natural Language Processing in Wikipedia texts. Proceedings XVIII GEOINFO. 2017: 122—127.
17. *Qiu Q., Xie Z., Wu L., Tao L.* Dictionary-based automated information extraction from geological documents using a Deep Learning Algorithm. Earth and Space Science. 2020, 7: 1—18. doi: 10.1029/2019EA000993.
18. *Wang C., Ma X., Chen J., Chen J.* Information extraction and knowledge graph construction from geoscience literature. Computers and Geosciences. 2018, 112: 112—120. doi: 10.1016/j.cageo.2017.12.007.
19. MyStem. Available at: https://yandex.ru/dev/mystem (date of access: 28.02.2021).
20. Microsoft Visual Studio 2019. Available at: https://visualstudio.microsoft.com/ru (date of access: 28.02.2021).
21. Surfer. Voxler. Golden Software Inc. Available at: https://www.goldensoftware.com/products (date accessed: 28.02.2021).

Конфликт интересов: конфликт интересов отсутствует.

Статья поступила 23.03.2021
Принята к публикации 03.05.2021

Сведения об авторе

Краснобородько Олег Юрьевич, заведующий лабораторией промысловой океанологии Атлантического филиала ФГБНУ «ВНИРО» («АтлантНИРО»), sea@atlantniro.ru.

Information about author

Krasnoborodko Oleg Yurievich, Head of the Laboratory of Fishery Oceanology of the Atlantic Branch of “VNIRO” (“AtlantNIRO”), sea@atlantniro.ru.