Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем

О. А. Митрофанова [1], Р. В. Голубев [1], П. А. Гусяцкая [1], К. В. Макеев [1], Е. А. Плюснина [1], Д. Д. Сухан [1,2], А. В. Трошина [1], А. А. Уткина [1]

[1] Санкт-Петербургский государственный университет, [2] Just AI

o.mitrofanova@spbu.ru, st110682@student.spbu.ru, st068584@student.spbu.ru, st110200@student.spbu.ru, st109958@student.spbu.ru, st110829@student.spbu.ru, st110338@student.spbu.ru, st110578@student.spbu.ru

Аннотация

В статье представлены результаты экспериментов по обучению семейства тематических моделей корпуса текстов по корпусной лингвистике, создаваемого на кафедре математической лингвистики СПбГУ под руководством В. П. Захарова. Тематическое моделирование корпуса ТКиКЛ осуществлено с помощью алгоритмов NMF, LSA, LDA, Biterm. Обобщение тем с помощью меток реализовано на основе обработки данных из выдачи информационно-поисковой системы, статических предсказывающих моделей Word2Vec, обученных на корпусе, а также большой языковой модели ChatGPT. Результаты тематического моделирования с назначением меток тем сопоставляются с данными о распределении докладов по тематическим блокам конференций в соответствии со схемой экспертной разметки.

Ключевые слова: корпусная лингвистика, материалы конференций, тематическое моделирование, метки тем, рубрикация

Библиографическая ссылка: Митрофанова О. А., Голубев Р. В., Гусяцкая П. А., Макеев К. В., Плюснина Е. А., Сухан Д. Д., Трошина А. В., Уткина А. А. Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 30–44. DOI: 10.17586/2541-9781-2024-8-30-44.

1. Введение

Данная статья посвящена решению задачи построения семейства тематических моделей корпуса текстов статей по корпусной лингвистике (далее корпус ТКиКЛ), составленного под руководством основателя Петербургской школы корпусной и компьютерной лингвистики Виктора Павловича Захарова. Разрабатываемый коллективом исследователей корпус представляет собой ценный источник информации о становлении и развитии методологии, ресурсов, понятийного аппарата и терминологии корпусной лингвистики. В статье [1] представлено описание процедуры формирования корпуса, разметки ключевых выражений в корпусе, генерации аннотаций и систематизации именованных сущностей. В статье [2] описаны эксперименты по формированию базы данных с метаинформацией и по разработке системы визуализации результатов поиска. Для улучшения качества поиска было проведено тематическое моделирование, результаты которого сопоставлены с ручной разметкой рубрик.

Под тематическим моделированием традиционно понимается особый способ построения структурно-семантической модели корпуса текстов, которая определяет взаимосвязи тем, документов и слов-тематизаторов [3]. Темы рассматриваются как скрытые факторы, представленные кластерами слов-тематизаторов. Каждый документ связан с одной или несколькими темами с некоторой вероятностью, при этом темы могут пересекаться. Наиболее распространенные методы тематического моделирования включают группу алгебраических моделей, например, латентный семантический анализ (Latent Semantic Analysis, LSA), неотрицательная матричная факторизация (Non-negative Matrix Factorization, NMF) и др., группу вероятностных моделей, например, вероятностный латентный семантический анализ (probabilistic Latent Semantic Analysis, LSA), латентный Распределение Дирихле (Latent Dirichlet Allocation, LDA) и т. д. В практических задачах широко используются мультимодальные версии тематических моделей, учитывающие дополнительные параметры корпусов (авторство текстов, время создания документов в корпусе, иерархия тем и т. д.), комбинируемые с моделями распределенных векторных вложений, например, BERTopic.

В статье представлены результаты построения тематических моделей корпуса ТКиКЛ с помощью алгоритмов NMF, LSA, LDA и Biterm. Параметры экспериментов были сходными, что обеспечивает объективность описания и сопоставления полученных результатов. Во всех экспериментах мы придерживались следующей схемы: проведение предварительной разметки n-грамм с помощью алгоритма выделения ключевых выражений RAKE [4], построение серии моделей с различным числом тем (5, 10, 15, 20), число слов-тематизаторов в темах (10, 15 20). В примерах, приводимых далее в статье, сохранено форматирование выдачи тематических моделей, предполагающее декапитализацию и отдельные случаи сохранения текстов в нелемматизированном варианте. Для оценки качества и интерпретируемости полученных моделей были определены значения агрегированной когерентности [5], перплексии [6] и энтропии [7]. Отдельные случаи расширения схемы экспериментов оговариваются в соответствующих разделах.

Тематические модели обеспечивают предпосылки для разведочного поиска в корпусах текстов. Повышение интерпретируемости моделей должно способствовать улучшению качества извлечения информации их текстов. Одним из факторов, влияющих на интерпретируемость тематических моделей, их адекватность решаемым задачам и исходным данным, является возможность обобщения тем с помощью меток [8; 9; 10]. Метка темы — это слово или словосочетание, отражающее общее содержание темы. Согласно традиции, темы условно обозначаются с помощью номера и первого слова-тематизатора, которое далеко не всегда является самым общим или типичным относительно темы. В автоматическом понимании текста разработаны формальные методы назначения меток тем, различающиеся источниками меток (внешними по отношению к корпусу и внутренними, использующими информацию из целевого корпуса), структурой меток (униграммы, биграммы, триграммы и т. д., выделяющиеся по лексико-грамматическим шаблонам), типами используемых алгоритмов. В [11] представлена апробация методов назначения меток тем на основе ИПС, с применением предсказаний дистрибутивно-семантических моделей и больших языковых моделей ChatGPT. Руководствуясь тем, что в экспериментах с корпусом научных новостных сообщений данный набор методов хорошо зарекомендовал себя, было принято решение воспроизвести его в проекте по тематическому моделированию корпуса ТКиКЛ.

2. Результаты тематического моделирования корпуса ТКиКЛ

2.1. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма NMF

Алгоритм NMF (Non-Negative Matrix Factorization, неотрицательная матричная факторизация) [12; 13] заключается в поиске для некой неотрицательной матрицы X двух

матриц (W, H), чье произведение будет являться приближением оригинальной матрицы X. В контексте тематического моделирования текстовых данных это означает, что для исходного корпуса подбираются матрицы «слова — темы» и «темы — документы», показывающие, соответственно, какие слова характеризуют каждую из тем, и как темы распределены по документам. К преимуществам алгоритма NMF относят высокую интерпретируемость результатов, вытекающую из неотрицательности элементов матриц, а также способность выявлять в данных более редкие и специфичные темы.

В настоящем проекте алгоритм NMF был применен к текстовым данным: каждая тема, таким образом, представляет из себя ранжированный по весам список слов и словосочетаний. Реализация алгоритма тематического моделирования NMF была осуществлена при помощи библиотеки scikit-learn [14]. Ход экспериментов предполагал предварительную разметку в корпусе униграмм и биграмм на основе алгоритма RAKE. Все слова, выделенные в составе биграмм, были затем удалены из неразмеченных текстов корпуса на этапе формирования списка уникальных униграмм. Биграммы были затем лемматизированы и представлены в корпусе в виде «практический_применение» — чтобы при обучении тематической модели для каждой биграммы формировался отдельный вектор. Список объединенных и лемматизированных биграмм был объединен со списком лемматизированных униграмм.

Далее была проведена серия экспериментов, целью которой было установить оптимальное количество тем, которые будет выделять модель NMF на корпусных данных. Для этого была проведена оценка когерентности для четырех моделей, выделивших 5, 10, 15 и 20 тем, соответственно. Наивысший показатель когерентности в группе моделей (≈ 0,44) был достигнут при 20 темах – это количество тем и было принято как рабочее значение параметра в финальной модели NMF. Данная модель была построена на корпусных данных, векторизованных при помощи метрики TF-IDF, в результате чего было получено 20 тем, представляющих собой отранжированные списки тематизаторов — униграмм и биграмм.

Результирующие темы демонстрируют не только высокий показатель когерентности, но и представляются достаточно интерпретируемыми при экспертной оценке. К примеру, в модели четко противопоставлены темы «Перевод» (*перевод, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо-, переводчик, многоязычный, словарь…*), «Коммуникативные стратегии» (*жест, ребенок, коммуникативный, движение, робот, мимика, поведение, эмоциональный, детский, рука, участник, человек, невербальный, социальный, собеседник, коммуникация, возраст, функция, мультимодальный, действие…*), «Медиапространство» (*театр, театральный, спектакль, сцена, зритель, интернет, новый, трансляция, александринский, режиссер, пространство, многопотоковый, видео, творческий, виртуальный, технология, театра, интерактивный, медиа, актер…*) и т.д.

Отдельно стоит упомянуть, что модель NMF успешно выделила в корпусе редкие темы, то есть темы, представленные в корпусе лексикой низкой частотности, к примеру тема «Финно-угорские языки» (*финский, ижорский, диалектный, диалект, песня, народный, прибалтийско-, карельский, говор, фонетический, вепсский, ингерманландия, топоним, топонимический, язык, текст, приток, топонимов, песен, звуковой…*), или «Тибетский язык» (*тибетский, разметка, грамматический, композит, тэг, токен, лексический, корпус, аффикс, буддийский, индийский, термин, разметить, традиция, сегментация, проект, трактат, традиции…*).

Заметим, что результаты применения алгоритма NMF на настоящий момент следует считать ориентиром в построении тематических моделей корпуса ТКиКЛ.

2.2. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма LSA

Латентный семантический анализ (LSA, Latent Semantic Analysis) — классический алгоритм построения дистрибутивных моделей корпусов текстов, основанный на матрично-векторных преобразованиях и отражающий близость значений и совместную встречаемость слов в корпусе [15; 16; 17]. Принцип работы LSA можно разбить на несколько этапов. На первом шаге текст предобрабатывается, затем токенам назначают веса (например, с помощью TF-IDF), и по этим весам строится матрица. В данном проекте для предобработки использовалась функция CountVectorizer в библиотеке scikit-learn [14]. Она токенизирует текст, после чего производит расчет вхождений, каждому токену присваивается уникальный целочисленный индекс, и эта информация приводится в формат матрицы. На финальной ступени матрица раскладывается методом сингулярного разложения (SVD, Singular Value Decomposition).

В экспериментах по обучению моделей LSA, согласно стандартной схеме, лучшие результаты были получены при выборе 15 тем. В этом случае темы четче разграничиваются, но при этом остаются довольно специализированными. Ниже приведены примеры общих тем: «Корпус как явление» (*текст, корпус, слово, являться, язык, система, русский, работа, данные, анализ, семантический, значение, информация, словарь, результат, иметь, использование, использовать, исследование, информационный…*); «Модель языка» (*слово, понятие, модель, термин, связь, язык, знание, отношение, электронный, семантический, корпус, развитие, определение, область, услуга, государственный, слов, поле, метафора, информационный…*); к более частным следует отнести лингвистические темы: «Семантика» (*слово, семантический, значение, ударение, иметь, глагол, понятие, класс, связь, отношение, стих, являться, предлог, объём, строка, определение, слов, модель, вид, часть…*); «Теория поэзии» (*ударение, объём, стих, строка, текст, слоговой, слог, слово, ударный, метр, икт, пропуск, место, электронный, показатель, интервал, схема, объёмный…*); прикладные лингвистические темы «Социальная сеть» (*социальный, сеть, пользователь, интернет, политический, сети, новый, сетевой, пространство, являться, человек, сми, текст, связь, коммуникация, исследование, аудитория, медиа, количество, сервис…*); «Электронное голосование» (*голосование, система, электронный, избиратель, голос, голосования, выборы, избирательный, список, интернет, проблема, слово, цифровой, ключ, бюллетень, возможность, дистанционный, помощь, кандидат, использовать…*).

Полученные темы органично вписываются в спектр исследовательских направлений, представленных в корпусе ТКиКЛ, однако для повышения интерпретируемости результатов модель LSA требует более точной настройки.

2.3. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма LDA

Скрытое распределение Дирихле (Latent Dirichlet allocation, LDA) — это широко используемый алгоритм вероятностного тематического моделирования, рассматривающий процесс определения тематической структуры текстов на основе семейства непрерывных многомерных вероятностных распределений [18]. Как известно, тематическая модель LDA частично решает проблему переобучения pLSA [19].

В экспериментах использовалась реализация алгоритма LDA в библиотеках scikit-learn [14] и genism [20]. Предобработка корпуса предполагала разметку в корпусе ключевых выражений (биграмм и триграмм) посредством алгоритма RAKE. Результаты обучения тематических моделей в библиотеках scikit-learn и genism отличаются долей неоднословных тематизаторов: например, LDA в scikit-learn в основном выделяет тематизаторы-униграммы, и единственной биграммой оказалось словосочетание «*социальная_сеть*», в то время как LDA в gensim генерирует темы с высокой долей биграмм и триграмм. С точки зрения интерпретируемости тем и равномерности распределения тем по документам следует отдать предпочтение варианту реализации LDA в библиотеке scikit-learn. Было проведено обучение серии моделей со сменой параметров (5, 10, 15 и 20 тем) и оценкой

когерентности. В результате экспериментов было установлено, что оптимальное число интерпретируемых наборов слов для LDA стремится к двадцати.

Среди полученных тем есть ядерные темы общего содержания, связанные с общей проблематикой корпуса ТКиКЛ, в частности, «Моделирование естественного языка» (*модель, алгоритм, формула, критерий, подобный, список, пример, метод, часть, параметр, следующий, ошибка, использовать…*), «Корпус текстов» (*корпус, словарь, русский, исследование, база, поиск, термин, материал, разметка, лингвистический, картотека, дескриптор, словоформа, анализ, лингвистика…*), «Представление текстов в корпусе» (*корпус, буква, словоформа, контекст, русский, написание, житие, вариант, словарь, рукопись, век, семантический, первый, термин, разметка…*). Примерно четверть сгенерированных тем соотносится с задачами семантического анализа, например: «Семантическая разметка» (*семантический, разметка, отношение, корпус, значение, разный, связь, признак, лингвистика, анализ, система, лексика, тип, общий…)*, «Формальные онтологии» (*онтология, корпус, понятие, отношение, возможность, класс, рамка, описание, элемент, слот, использование, экземпляр, система, иерархия, онторедактор…*), «Семантические отношения» (*отношение, семантический, словарь, синсет, значение, лексический, структура, существительное, связь, система, лексико-, глагол, база, часть, словоформа…*) и некоторые другие. Наряду с этим, одиночные темы представляют такие специфичные направления компьютерной и корпусной лингвистики, как «Морфосинтаксическая разметка» (*предложение, связь, оценка, корпус, синтаксический, парсер, узел, структура, отношение, этап, морфологический, речь, тип, случай, часть…*), «Звуковые корпусы текстов» (*речь, эда, материал, русский, устный, рассказ, речевой, корпус, живой, языковой, составлять, звуковой, вариант, запятая, точка…*).

2.4. Тематическая модель корпуса ТКиКЛ, построенная с помощью алгоритма Biterm

Модель битермов (Biterm Topic Model) [21] создана для распознавания тем в коротких текстах, таких как твиты и посты социальных сетей. Модели типа LSA или LDA недостаточно приспособлены для обработки текстов данного типа, поскольку неявно учитывают совместную встречаемость слов в документах, что проявляется в виде тематической разреженности данных в коротких текстах. BTM генерирует темы, напрямую моделируя шаблоны битермов (биграмм) для всего корпуса текстов. Восстановление битермов по шаблонам улучшает качество тем, а агрегированные шаблоны битермов для корпуса в целом решают проблему тематической разреженности.

BTM справляется с задачей тематического моделирования благодаря сэмплированию по Гиббсу. Основная идея сэмплирования заключается в генерации образцов из совместного распределения вероятностей путём итеративной выборки из условных распределений каждой переменной с учётом значений всех остальных переменных. Этот процесс позволяет получать выборки из сложных, высокоразмерных распределений, разбивая задачу на более простые условные шаги выборки. Сэмплирование по Гиббсу — это разновидность метода Монте-Карло с цепью Маркова, широко используемого в байесовской статистике и машинном обучении для аппроксимации апостериорных распределений.

Для экспериментов использовалась библиотека bitermplus [22]. На входе модели, помимо корпуса текстов нужно подать предполагаемое число тем. В ходе эксперимента были протестированы 5 значений: 5, 8, 10, 15 и 20 тем (здесь к стандартной схеме было добавлено еще одно значение — 8 тем). Примеры результирующих тем приведены ниже: «Информационные технологии» (*информационный, электронный, система, государственный, развитие, являться, научный, информация, работа, использование…*), «Корпус» (*текст, корпус язык, работа, словарь, система, анализ, данные, русский, разметка…*), «Семантика» (*слово, значение, семантический, являться, глагол, текст, форма, тип, случай, два…*) и т.д. Наибольшая когерентность наблюдается у модели с пятью темами, однако перплексия и энтропия у данной модели выше, чем у остальных. Средняя

когерентность почти не меняется с возрастанием количества тем, а энтропия даже падает. Результаты показывают, что оптимальное число тем определяется в промежутке между 16 и 20 темами.

3. Результаты генерации меток тем в корпусе ТКиКЛ

3.1. Генерация меток тем для текста с помощью ИПС

Одним из вариантов генерации меток тем является применение информационно-поисковых систем (ИПС). ИПС и тематическое моделирование тесно связаны, однако обычно именно метки тем используются для улучшения веб-поиска, обратная же схема встречается редко. Поскольку в основе всех современных ИПС лежит принцип отбора наиболее релевантных запросу документов, можно использовать это свойство для решения поставленной задачи. Так как тексты веб-документов могут быть разного объёма и содержать разнородную информацию, удобнее реализовать схему, при которой результаты выдачи используются не полностью, а частично. Можно рассмотреть два варианта — суммаризацию всего документа, например, с помощью моделей-трансформеров, либо использование только заголовков веб-страниц без учета основной части текста. Может показаться, что вторая опция ограничивает наши возможности, однако она более выгодна, так как при оценке релевантности ИПС учитывает заголовки с большим весом. Кроме того, это обеспечивает бОльшую вероятность получения связного текста, в отличие от автоматической суммаризации.

Методика генерации меток тем с помощью ИПС, примененная в нашем исследовании, является модификацией метода, представленного в [8; 11], и так же, как и исходный вариант, состоит из нескольких этапов.

На первом этапе в качестве входных данных используются списки тем, сгенерированных исследуемыми алгоритмами тематического моделирования. Для каждой метки тем отправляется запрос в ИПС Google [23] для получения поисковой выдачи. При этом все метки рассматриваются как единое предложение, как и в случае обычных запросов в ИПС, которые не всегда характеризуются синтаксической связностью. Использование Google обусловлено тем, что эта поисковая система не блокирует последовательные автоматические запросы к своему API, в отличие от Yandex. Полученная выдача далее фильтруется, рассматриваются 30 первых по релевантности документов.

На втором этапе для всех заголовков темы составляется матрица совместной встречаемости слов в контекстном окне [-1, 1], что позволяет выделить биграммы. Такую матрицу можно визуализировать как взвешенный граф, где рёбра — это связи в биграммах, а веса — встречаемость. Для слов, не встречавшихся друг с другом, устанавливается минимальный вес, равный 1.

Третий этап определяется как Power Iteration или применение степенного метода. Он используется также и в PageRank, алгоритме, имеющем ключевое значение в современных ИПС. Из матрицы, составленной на предыдущем этапе, собирается стартовое состояние: это словарь, где ключи — это все слова, а значения равны величине, обратной количеству слов. Затем в матрице совместной встречаемости все значения делятся на сумму элементов в этой строке. Наконец, запускается алгоритм сходимости, в ходе которой предыдущее стартовое состояние заменяется скалярным произведением предыдущего словаря на матрицу совместной встречаемости. Это происходит либо фиксированное число раз (в нашем алгоритме — 1000), либо пока разница между предыдущим и новым состоянием не становится меньше некоторого числа ε. Иначе говоря, рассчитывается собственный вектор для матрицы. Затем все слова сортируются по весу и из них отбирается n наиболее вероятных.

Два финальных этапа представляют собой формирование и фильтрацию полученных меток на основе правил. Для набора заголовков темы составляется список n-грамм.

Биграммы и триграммы составляются по правилам, а большие — из меньших по принципу паззла (конец одной биграммы равен началу предыдущей). Таким образом, максимальный размер n-грамм — 6 токенов. Правила составления n-грамм направлены на формирование наиболее частотных для русского языка словосочетаний — например, ADJ + NOUN или NOUN + NOUN (GEN, INSTR). Каждая n-грамма получает вес, равный сумме весов её составляющих, после чего отбираются первые 5 n-граммов по весу.

Постобработка включает фильтрацию n-грамм по следующим правилам: удаление повторов и совпадений; удаление меток с повторениями одного и того же слова; коррекция согласования внутри словосочетаний; отсеивание слишком коротких слов или случайных букв. На выходе метод производит от одной до трех биграмм для каждой темы, как правило, являющиеся осмысленными словосочетаниями. Примеры результирующих меток представлены в таблице 1.

Таблица 1. Примеры меток тем, сгенерированных ИПС

Темы (фрагмент выдачи)	**Метки ИПС**
онтология, понятие, свойство, отношение, знание, термин, связь, сущность, определение, объект, онтологии, класс, предметный, смысл, система, модель, являться, граф, множество, семантический…	*знания и онтология, онтология и тезаурусы, подход к процессам и системы*
мера, биграмм, коллокация, частота, словосочетание, коллокат, статистический, слово, сочетание, встречаемость, список, мер, словосочетаний, сочетаемость, значение, коллокаций, оценка, результат, корпус, эксперимент…	*выделение коллокаций, значение коллокаций, образование и наука*
учебный, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа…	*студенты и преподаватели в процессе, традиции к инновациям в обучении, инновации в обучении*
перевод, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо, переводчик, многоязычный, словарь…	*перевод для выравнивания, переводчик и редактор, поиск и ранжирование*
житие, текст, цитата, агиографический, житийный, рукопись, скат, разметка, написание, текста, древнерусский, рукописный, словоуказатель, издание, рукописи, фрагмент, дионисий, алексеева, глушицкого, представление…	*разметка в корпус агиографический текст, корпус агиографический текст, представление и анализ элементов структуры*
государственный, электронный, орган, гражданин, информационный, услуга, развитие, власть, социальный, правительство…	*информация о орган государственный власть, министерство социальный политика и труд, услуги для граждан и бизнес*
модель, алгоритм, тематический, слово, метод, текст, документ, тема, оценка, слов, результат, анализ, задача, распределение, количество, эксперимент, матрица, коллекция, вероятность, вероятностный…	*качество тематический модель для задача, плотность многомерных распределений в виде, методология и методы научных исследований*
…	*…*

Среди преимуществ использования ИПС для генерации меток тем следует отметить возможность генерации меток различной длины, более высокий уровень согласованности и объективности итоговых меток благодаря отбору релевантных комбинаций, а также достаточно высокую скорость исполнения. Метод способен продуцировать длинные осмысленные сочетания, такие как «типология ассоциативных словарей русского языка» или «основы цифровой грамотности и кибербезопасность». Кроме того, применение

поисковых методик понижает нестабильность некоторых базовых моделей, прежде всего, LDA.

Однако применение ИПС для тематического моделирования обладает и недостатками. В их числе сложная для настройки структура, построенная на разных принципах. Результаты не детерминированы, как и в случае нейросетей, и зависят от результатов работы поисковых систем, которые периодически обновляются. Результат зависит и от выбранной методики тематического моделирования: в данном исследовании лучший результат был получен для моделей LSA и NMF, где метки оказались более интерпретируемы. Метод сложно заставить производить фиксированное количество меток, а в некоторых случаях он может не создать ни одной. Поэтому рекомендуется использовать ИПС наряду с другими методами для получения лучшего результата.

3.2. Генерация меток тем для текста с помощью дистрибутивно-семантических моделей Word2Vec

Данный способ генерации меток тем относится к числу подходов, позволяющих назначать метки тем на основе внутренних по отношению к корпусу источников. Как предлагается в [9; 10; 11], мы применяли статические дистрибутивно-семантические модели типа Word2Vec [24] и рассматривали их предсказания как кандидаты в метки тем. Нейросетевая архитектура Word2Vec представляет контексты корпуса в виде векторов, которые при условии близости значения и употребления слов локализуются сходным образом, о чем свидетельствуют высокие значения косинусной меры. В Word2Vec предсказание близких лексических единиц осуществляется с помощью функции most_similar, допускающей генерацию ассоциатов как для отдельного слова, так и для группы слов, в нашем случае представляющей собой набор слов-тематизаторов, представляющих отдельную тему. Для предсказания кандидатов в метки тем на предобработанном корпусе ТКиКЛ были обучены две модели CBOW (Continuous Bag of Words) и Skip-gram, которые по-разному фиксируют отношения между словами в модели: если модель CBOW предсказывает потенциальные замены целевого слова с учетом контекста, то модель Skip-gram позволяет предсказывать элементы контекстного окружения для целевого слова. Для корректного обучения моделей корпус ТКиКЛ был токенизирован и повторно лемматизирован, для исключения попадания служебных слов и иных незнаменательных элементов в метках тем был подключен стоп-словарь. Частеречная разметка корпуса для того, чтобы исключить попадание наречий, прилагательных и иных частей речи кроме существительных в метки тем. Результаты экспериментов дают основания для дискуссии о статусе сгенерированных меток, которые действительно уточняют содержание тем, однако не обобщая их, а скорее расширяя. Примеры меток CBOW и Skip-gram представлены в таблице 2. Совпадения предсказаний двух типов моделей указывают на то, что повторяющиеся кандидаты в метки (выделены полужирным) являются релевантными для тем. Модели Word2Vec, обученные на корпусе с предварительной разметкой ключевых выражений с помощью алгоритма RAKE, генерируют повторяющиеся метки, что следовало бы избежать. Наиболее интерпретируемые результаты были получены в комбинации моделей Word2Vec и тем, порожденных моделями LSA и NMF.

Таблица 2. Примеры меток тем, сгенерированных моделями Word2Vec

Темы (фрагмент выдачи)	CBOW	Skip-gram
научный, система, информационный, библиотека, сервис, пользователь, поиск, ресурс, электронный, данные, информация, полнотекстовый, база, поддержка, программный, проект, запрос, поисковый, публикация, доступ…	*интерфейс, карта, контент, пользовательский*	*вебсайт, протокол, ипс, навигация*

Продолжение таблицы 2

Темы (фрагмент выдачи)	CBOW	Skip-gram
онтология, понятие, свойство, отношение, знание, термин, связь, сущность, определение, объект, онтологии, класс, предметный, смысл, система, модель, являться, граф, множество, семантический…	*иерархия, именовать, элементарный, родовидовой*	*таксономия, экземпляр, иерархия, родовидовой*
словарь, русский, словарный, слово, языка, словаря, язык, лексикографический, корпус, толковый, картотека, грамматический, словарей, академический, цитата, бас, слова, новый, создание, рнк…	*словник, двуязычный, одноязычный, зализняк*	*указатель, словник, грамматик, бумажный*
морфологический, синтаксический, разметка, слово, грамматический, форма, предложение, словоформа, разбор, существительное, неоднозначность, автоматический, ошибка, парсер, омонимия, часть, анализатор, снятие, падеж, вариант…	*частеречной, тег, помета, лемматизация*	*частичный, частеречной, морфема, лемматизация*
учебный, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа…	*перспектива, филология, методический, методология*	*учитель, будущий, привлечение, методический*
семантический, глагол, значение, валентность, слово, лексический, контекст, синсет, лексико-, актант, употребление, иметь, описание, отношение, семантика, класс, единица, синтаксический, языковой…	*ядро, синонимия, стилистический, номинация*	*сосед, синонимия, корень, синтаксема*
модель, алгоритм, тематический, слово, метод, текст, документ, тема, оценка, слов, результат, анализ, задача, распределение, количество, эксперимент, матрица, коллекция, вероятность, вероятностный…	*статистика, гипотеза, ранжирование, классификатор*	*lda, отзыв, классификатор, кластеризация*
…	…	…

3.3. Генерация меток тем для текста с помощью большой языковой модели ChatGPT

Проводимое исследование открывает новые возможности в тестировании больших языковых моделей, в частности, мощной языковой модели ChatGPT, созданной OpenAI. В эксперименте использовалась модель GPT-3.5 [26]. Модель способна генерировать текст, имитируя стиль носителя языка и понимая контекст. Далее рассмотрим применение ChatGPT к генерации меток с учетом весов слов (в этом состоит модификация протокола, реализованного в сходных задачах [11; 25]).

В ходе эксперимента на вход модели подавались темы в виде наборов слов-тематизаторов с их весами. При обращении к модели использовались промпты, например, «Используя слова из списка, составь несколько общих выражений и выдели главное слово». Предсказанные кандидаты в метки тем и ключевые слова были сохранены в таблицах Excel. ChatGPT успешно выделял общие выражения, адекватно отражающие тематику текстов корпуса ТКиКЛ, однако в ходе анализа данных были выявлены некоторые особенности. Во-первых, ChatGPT при использовании одного и того же чата ChatGPT может запоминать структуру диалога и тем самым в предсказаниях опирается на излишне широкий контекст, что приводит к семантическим сдвигам в генерации меток тем. Во-вторых, веса слов-

тематизаторов оказывают влияние на процесс генерации, что может привести к искажению результатов. В-третьих, в промпте желательно явно указывать ожидаемое количество кандидатов в метки для получения реалистичных результатов.

ChatGPT демонстрирует высокий потенциал для генерации меток, но требует аккуратного использования. Память контекста и влияние весов могут сказаться на точности результата. Верификация полученных меток была проведена с привлечением схемы рубрикации, содержащей темы работы секций конференций [1]. Рубрики из схемы экспертной рубрикации приведены в последней колонке таблицы 3.

Таблица 3. Примеры меток тем, сгенерированных ChatGPT

Темы (фрагмент выдачи)	**Метки ChatGPT**	**Главное слово в теме, выбор ChatGPT**	**Рубрика из схемы экспертной рубрикации**
научный, система, информационный, библиотека, сервис, пользователь, поиск, ресурс, электронный, данные, ***информация****, полнотекстовый, база, поддержка, программный, проект, запрос, поисковый, публикация, доступ*	*«Информационные ресурсы в научных исследованиях», «Электронные библиотеки и доступ к информации», «Сервисы для поиска научных данных»*	*информация*	Создание, разработка и применения корпусов
мера*, биграмм, коллокация, частота, словосочетание, коллокат, статистический, слово, сочетание, встречаемость, список, мер, словосочетаний, сочетаемость, значение, коллокаций, оценка, результат, корпус, эксперимент*	*«Статистический анализ коллокаций в корпусе», «Оценка частоты и сочетаемости слов», «Меры и статистика в анализе словосочетаний»*	*мера*	Статистические исследования на материале корпусов
словарь*, русский, словарный, слово, языка, словаря, язык, лексикографический, корпус, толковый, картотека, грамматический, словарей, академический, цитата, бас, слова, новый, создание, рнк*	*«Лексикографический анализ русского языка», «Словарные ресурсы и их создание», «Грамматический аспект в словарях»*	*словарь*	Корпусы и лексикография
морфологический, синтаксический, ***разметка****, слово, грамматический, форма, предложение, словоформа, разбор, существительное, неоднозначность, автоматический, ошибка, парсер, омонимия, часть, анализатор, снятие, падеж, вариант*	*«Морфологическая и синтаксическая разметка», «Грамматический анализ предложений», «Автоматический разбор текста и ошибки»*	*разметка*	Морфология и синтаксис в корпусах
учебный*, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология, материал, использование, тест, профессиональный, возможность, работа*	*«Обучение и образование с использованием корпусов», «Английский язык в образовательной среде», «Корпусный анализ ошибок студентов»*	*учебный*	Обучающие корпусы

Продолжение таблицы 3

Темы (фрагмент выдачи)	Метки ChatGPT	Главное слово в теме, выбор ChatGPT	Рубрика из схемы экспертной рубрикации
***перевод**, параллельный, текст, переводный, английский, выравнивание, машинный, система, предложение, двуязычный, перевода, термин, корпус, создание, терминологический, текстов, англо, переводчик, многоязычный, словарь*	*«Перевод и анализ текстов в параллельном корпусе», «Машинный перевод и создание терминологических словарей», «Текстовый анализ и создание переводческих ресурсов»*	*перевод*	Параллельные корпусы и машинный перевод
***житие**, текст, цитата, агиографический, житийный, рукопись, скат, разметка, написание, текста, древнерусский, рукописный, словоуказатель, издание, рукописи, фрагмент, дионисий, алексеева, глушицкого, представление*	*«Житийные тексты и их агиографические особенности», «Написание и разметка древнерусских рукописей», «Текстологический анализ житийных материалов»*	*житие*	Исторические корпусы
…	…	…	…

4. Заключение

Разработка специализированных корпусов текстов, к которым относится корпус ТКиКЛ, созданный коллективом авторов под руководством В. П. Захарова, требует как тщательной подготовки текстов, так и создания инструментария для автоматизации извлечения, и структурирования информации в корпусе. По этой причине столь важно успешное проведение экспериментов по тематическому моделированию корпуса ТКиКЛ, которое показало особенности структурно-семантической и тематической организации корпуса. Важно отметить, что в построении тематических моделей корпуса авторы следовали принципу мультимодальности и учитывали возможность совмещения базового протокола тематического моделирования с автоматическим выделением ключевых выражений и автоматическим назначением меток тем. Наиболее интерпретируемыми оказались результаты, полученные с помощью алгоритма NMF с метками тем, сгенерированными с помощью ChatGPT. Объективность полученных результатов подтверждается соответствием между автоматически назначенными метками и рубриками их схемы экспертной разметки, составленной на основе программ работы конференций, материалы которых представлены в корпусе ТКиКЛ.

Литература

[1] Митрофанова О. А., Адамова М. А., Букреева Л. А., Зернова А. К., Литвинова А. А., Павликова В. С., Сологуб П. С. Корпус текстов по корпусной лингвистике: состав и этапы формирования // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 12–28. DOI: 10.17586/2541-9781-2024-8-12-28.

[2] Сухан Д. Д., Плюснина Е. А. Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии.

Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 44–59. DOI: 10.17586/2541-9781-2024-8-44-59.

[3] Воронцов К. В. Вероятностное тематическое моделирование: Теория регуляризации ARTM и библиотека с открытым кодом BigARTM. М.: URSS, 2023. 208 с.

[4] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings. FRC CSC RAS. 2018. P. 369–372.

[5] Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. P. 262–272.

[6] Heinrich G. Parameter estimation for text analysis: Technical report. 2005. P. 1-32.

[7] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. 2018. Т. 512. P. 1192–1204.

[8] Ерофеева А., Митрофанова О. Автоматическое назначение меток тем в тематических моделях русскоязычных корпусов текстов // Структурная и прикладная лингвистика. Т. 12. СПб., 2019. С. 122–147.

[9] Kriukova A., Erofeeva A., Mitrofanova O., Sukharev K. Explicit Semantic Analysis as a Means for Topic Labeling // Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings. Springer, Cham. 2018. P. 167–177.

[10] Mitrofanova O., Kriukova A., Shulginov V., Shulginov V. E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science. Springer. 2021. Vol. 1357. P. 102–114.

[11] Mitrofanova O. A., Athugodage M. M., Ten L. V. Topic Label Generation in the Popular Science Corpus // Digital Geography: Proceedings of the International Conference on Internet and Modern Society (IMS 2023). Springer, 2023. (В печати)

[12] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. 2020. Vol. 12469, pt. 2. P. 134–152.

[13] Kuang D., Choo J., Park H. Nonnegative matrix factorization for interactive topic modeling and document clustering // Partitional clustering algorithms. 2015. P. 215–243.

[14] Scikit-learn // Scikit-learn. URL: https://scikit-learn.org/ (дата обращения: 09.02.2024).

[15] Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25 (2–3). P. 259–284.

[16] Чижик А. В. Использование методов тематического моделирования для оценки степени влияния СМИ на общественное настроение // Компьютерная лингвистика и вычислительные онтологии. Вып. 5. (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24–26 июня 2021 г. Сборник научных статей). СПб.: Университет ИТМО, 2021. С. 70–78.

[17] Кирина М. А. Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2022. Т. 20, № 2. С. 93–109.

[18] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. Vol. 3. P. 993–1022.

[19] Hofmann T. Probabilistic latent semantic indexing // ACM SIGIR Forum. 2017. Vol. 51 (2). P. 211–218.
[20] Gensim // Gensim. URL: https://radimrehurek.com/gensim/ (дата обращения: 09.02.2024).
[21] Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts // WWW 2013. Proceedings of the 22nd International Conference on World Wide Web. 2013. P. 1445–1456.
[22] biterm 0.1.5 // PyPI. URL: https://pypi.org/project/biterm/ (дата обращения: 09.02.2024).
[23] Google // Google. URL: https://www.google.ru/ (дата обращения: 09.02.2024).
[24] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 09.02.2024).
[25] Митрофанова О. А. Поиск и ранжирование текстов в специальном корпусе на основе тематического моделирования // Труды Международной конференции «Корпусная лингвистика — 2023», SPb Corpora 2023, 21–23 июня 2023 г. СПб.: Изд-во СПбГУ, 2024. (В печати)

Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment

O. A. Mitrofanova [1], R. V. Golubev [1], P. A. Gusyatskaya [1], K. V. Makeev [1], E. A. Pliusnina [1], D. D. Sukhan [1,2], A. V. Troshina [1], A. A. Utkina [1]

[1] Saint–Petersburg State University, [2] Just AI

The article presents novel experimental results concerning experiments aimed at training a family of topic models of the corpus on Corpus Linguistics, developed at the Department of Mathematical Linguistics, St. Petersburg State University under the supervision of V. P. Zakharov. Topic modelling of the corpus was carried out using NMF, LSA, LDA, Biterm algorithms. Ggeneralization of topics using labels is implemented on the basis of processing data from the output of an information search engine, static predictive Word2Vec models trained on the corpus, as well as a large ChatGPT language model. The results of topic modelling with the assignment of topic labels are compared with data on the distribution of reports by conference thematic blocks of in accordance with the expert markup scheme.

Keywords: corpus linguistics, conference materials, topic modelling, topic labels, rubrication

Reference for citation: Mitrofanova O. A., Golubev R. V., Gusyatskaya P. A., Makeev K. V., Pliusnina E. A., Sukhan D. D., Troshina A. V., Utkina A. A. Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 30–44. DOI: 10.17586/2541-9781-2024-8-30-44.

Reference

[1] Mitrofanova O. A., Adamova M. A., Bukreeva L. A., Zernova A. K., Litvinova A. A., Pavlikova V. S., Sologub P. S. Korpus tekstov po korpusnoj lingvistike: sostav i etapy formirovaniya // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXVII Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2024, Sankt-Peterburg, 24–26 iyunya 2024 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2024. S. 12-28. DOI: 10.17586/2541-9781-2024-8-12-28. (In Russian)

[2] Sukhan D. D., Pliusnina E. A. Metarazmetka i vizualizaciya dannyh v korpuse tekstov po korpusnoj lingvistike // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vypusk 8 (Trudy XXVII Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2024, Sankt-Peterburg, 24–26 iyunya 2024 g. Sbornik nauchnyh statej). — SPb.: Universitet ITMO, 2024. S. 44–59. DOI: 10.17586/2541-9781-2024-8-44-59. (In Russian)
[3] Vorontsov K. V. Veroyatnostnoe tematicheskoe modelirovanie: Teoriya regulyarizacii ARTM i biblioteka s otkrytym kodom BigARTM. M.: URSS, 2023. 208 s. (In Russian)
[4] Moskvina A., Sokolova E., Mitrofanova O. KeyPhrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings. FRC CSC RAS. 2018. P. 369–372.
[5] Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. P. 262–272.
[6] Heinrich G. Parameter estimation for text analysis: Technical report. 2005. P. 1–32.
[7] Koltcov S. Application of Rényi and Tsallis entropies to topic modeling optimization // Physica A: Statistical Mechanics and its Applications. 2018. T. 512. P. 1192–1204.
[8] Erofeeva A., Mitrofanova O. Avtomaticheskoe naznachenie metok tem v tematicheskih modelyah russkoyazychnyh korpusov tekstov // Structural and applied linguistics. Volume 12. St. Petersburg, 2019. P. 122–147. (In Russian)
[9] Kriukova A., Erofeeva A., Mitrofanova O., Sukharev K. Explicit Semantic Analysis as a Means for Topic Labeling // Artificial Intelligence and Natural Language Processing: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings. Springer, Cham. 2018. P. 167–177.
[10] Mitrofanova O., Kriukova A., Shulginov V., Shulginov V. E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science. Springer. 2021. Vol. 1357. P. 102–114.
[11] Mitrofanova O. A., Athugodage M. M., Ten L. V. Topic Label Generation in the Popular Science Corpus // Digital Geography: Proceedings of the International Conference on Internet and Modern Society (IMS 2023). Springer, 2023. (In print)
[12] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. 2020. Vol. 12469, pt. 2. P. 134–152.
[13] Kuang D., Choo J., Park H. Nonnegative matrix factorization for interactive topic modeling and document clustering // Partitional clustering algorithms. 2015. P. 215–243.
[14] Scikit-learn // Scikit-learn. URL: https://scikit-learn.org/ (access date: 09.02.2024).
[15] Landauer T. K., Foltz P. W., Laham D. Introduction to Latent Semantic Analysis // Discourse Processes. 1998. Vol. 25 (2–3). P. 259–284.
[16] Chizhik A. V. Ispol'zovanie metodov tematicheskogo modelirovaniya dlya ocenki stepeni vliyaniya SMI na obshchestvennoe nastroenie // Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vyp. 5. (Trudy XXIV Mezhdunarodnoj ob"edinennoj nauchnoj konferencii «Internet i sovremennoe obshchestvo», IMS-2021, Sankt-Peterburg, 24–26 iyunya 2021 g. Sbornik nauchnyh statej). SPb.: Universitet ITMO, 2021. S. 70–78. (In Russian)
[17] Kirina M. A. Sravnenie tematicheskih modelej na osnove LDA, STM i NMF dlya kachestvennogo analiza russkoj hudozhestvennoj prozy maloj formy // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul'turnaya kommunikaciya. 2022. T. 20, № 2. S. 93–109. (In Russian)

[18] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of machine Learning research. 2003. Vol. 3. P. 993–1022.
[19] Hofmann T. Probabilistic latent semantic indexing // ACM SIGIR Forum. 2017. Vol. 51 (2). P. 211–218.
[20] Gensim // Gensim. URL: https://radimrehurek.com/gensim/ (access date: 09.02.2024).
[21] Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts // WWW 2013. Proceedings of the 22nd International Conference on World Wide Web. 2013. P. 1445–1456.
[22] biterm 0.1.5 // PyPI. URL: https://pypi.org/project/biterm/ (access date: 09.02.2024).
[23] Google // Google. URL: https://www.google.ru/ (access date: 09.02.2024).
[24] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013. URL: https://arxiv.org/abs/1301.3781 (access date: 09.02.2024).
[25] Mitrofanova O. A. Poisk i ranzhirovanie tekstov v special'nom korpuse na osnove tematicheskogo modelirovaniya // Trudy Mezhdunarodnoj konferencii «Korpusnaya lingvistika — 2023», SPb Corpora 2023, 21–23 iyunya 2023 g. SPb.: Izd-vo SPbGU, 2024. (In Russian; in print)