Интеллектуальный анализ данных в корпусе текстов по корпусной и компьютерной лингвистике

О.А. Митрофанова, М.А. Адамова, Л.А. Букреева, Р.В. Голубев, П.А. Гусяцкая, А.К. Зернова, А.А. Литвинова, К.В. Макеев, В.С. Павликова, Е.П. Плюснина, П.Ю. Сологуб, Д.Д. Сухан, А.В. Трошина, А.А. Уткина

***Аннотация*— Статья посвящена описанию экспериментов, проводимых на материале корпуса статей по корпусной и компьютерной лингвистике, создаваемого на кафедре математической лингвистики СПбГУ. Корпус создан под руководством В.П. Захарова и включает в себя тексты докладов конференции «Корпусная лингвистика» с 2002 по 2021 гг., семинара «Компьютерная лингвистика и вычислительные онтологии» с 2011 по 2023 гг., а также некоторые другие материалы. В процессе разработки корпуса была проведена унификация формата представления текстов, исследована структура статей. Осуществлены эксперименты по генерации ключевых слов и аннотаций в тех случаях, когда авторский текст не содержал данную информацию. Исследованы типы именованных сущностей, зафиксированных в корпусе, реализован алгоритм их разметки. Проведен анализ распределения докладов по тематическим блокам конференций в соответствии со схемой экспертной разметки. Представлены результаты экспериментов по обучению семейства тематических моделей (NMF, LSA, LDA, Biterm) корпуса текстов. 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, компьютерная лингвистика, материалы конференций, разметка, ключевые слова, аннотации, тематическая разметка, именованные сущности, тематическое моделирование, метки тем, рубрикация.**

I. Введение

Проект, представленный в данной статье, посвящен памяти основателя и руководителя Петербургской школы корпусной и компьютерной лингвистики Виктора Павловича Захарова, нашего учителя и коллеги, который с 2002 года был главным организатором конференций и семинаров, где обсуждались проблемы создания и применения корпусов текстов. За двадцатилетний период проведения научных встреч были собраны ценные материалы, которые связаны с историей корпусной и компьютерной лингвистики, с развитием основных направлений, с кругом проблем и предлагаемых решений, с исследованием этапов становления и изменений терминологии рассматриваемой предметной области, ее логико-понятийной схемы и принципов стандартизации. Цель проекта состояла в разработке комплексного корпусного и терминологического ресурса с возможностью многопараметрического поиска источников. Результаты предшествующих исследований представлены в [1-3]. В данной статье рассматриваются следующие решенные нами задачи:

1) формирование корпуса статей по корпусной и компьютерной лингвистике (ТКиКЛ),
2) типы информации, представленные в корпусе,
3) разметка ключевых выражений в корпусе,
4) генерация аннотаций статей,
5) систематизация и разметка именованных сущностей,
6) мультимодальная тематическая разметка текстов корпуса на основе тематического моделирования с автоматическим назначением меток тем,
7) экспертная разметка рубрик в корпусе.

Помимо этих задач были решены задачи по формированию базы данных с метаинформацией и по

Статья получена 25 сентября 2024 г.

Митрофанова Ольга Александровна, Санкт-Петербургский государственный университет, канд. филол. наук, доцент (e-mail: o.mitrofanova@spbu.ru).

Адамова Мария Антоновна, Санкт-Петербургский государственный университет (e-mail: st110061@student.spbu.ru)

Букреева Людмила Александровна, Санкт-Петербургский государственный университет (e-mail: st110502@student.spbu.ru)

Голубев Ростислав Васильевич, Санкт-Петербургский государственный университет (e-mail: st110682@student.spbu.ru)

Гусяцкая Полина Андреевна, Санкт-Петербургский государственный университет (e-mail: st068584@student.spbu.ru)

Зернова Алиса Кирилловна, Санкт-Петербургский государственный университет (e-mail: st068103@student.spbu.ru)

Литвинова Анна Артемовна, Санкт-Петербургский государственный университет (e-mail: st110228@student.spbu.ru)

Макеев Кирилл Владимирович, Санкт-Петербургский государственный университет (e-mail: st110200@student.spbu.ru)

Павликова Владислава Станиславовна, Санкт-Петербургский государственный университет (e-mail: st109999@student.spbu.ru)

Плюснина Елизавета Алексеевна, Санкт-Петербургский государственный университет (e-mail: st109958@student.spbu.ru)

Сологуб Полина Юрьевна, Санкт-Петербургский государственный университет (e-mail: st095317@student.spbu.ru)

Сухан Даниил Дмитриевич, Санкт-Петербургский государственный университет (e-mail: st110829@student.spbu.ru)

Трошина Александра Валерьевна, Санкт-Петербургский государственный университет (e-mail: st110338@student.spbu.ru)

Уткина Александра Алексеевна, Санкт-Петербургский государственный университет (e-mail: st110578@student.spbu.ru)

Статья подготовлена по итогам выступления на Международной объединённой конференции «Интернет и современное общество» (IMS-2024).

разработке системы визуализации результатов поиска.

II. Состав и структура корпуса текстов ТКиКЛ

Процедура формирования корпуса ТКиКЛ на основе материалов конференций *Corpora* и *IMS CompLing* была комплексной, соответствовала протоколу, описанному в [4], и включала в себя несколько этапов.

Первый этап предполагал формирование электронной коллекции из текстов, опубликованных в материалах трудов конференций, которые включают статьи и тезисы (список изданий и их количественные параметры представлены в табл. 1). Тексты без аннотаций, авторских наборов ключевых слов и ссылок были преобразованы в файлы формата **.txt* для дальнейшей обработки. Названия файлов были стандартизированы: в них обязательно входит название конференции (*Corpora*/*IMS*) и год публикации сборника. Статьи на английском языке не были включены в корпус и не извлекались из сборников.

На втором этапе разработки корпуса был разработан и применен код на языке программирования Python, который корректировал имена файлов для обеспечения единообразия, а также проводил лемматизацию всех файлов в папках разных годов с помощью библиотеки Pymorphy2, что дало дополнительное деление корпуса на тексты с без лемматизации и с лемматизацией (*Corpora_raw* / *Corpora_lemmatized*, *IMS_raw* / *IMS_lemmatized*). Удаление нетекстовых элементов и лемматизация способствуют повышению качества анализа содержания текста и получить более точные результаты при использовании инструментов автоматической обработки текста.

Третий этап формирования корпуса состоял в сборке массива лемматизированных файлов (*full_corpus*) для дальнейшей их обработки, включающей следующие этапы: 1) автоматическое выделение ключевых слов и выражений, 2) автоматическая генерация аннотаций, 3) тематическое моделирование, 4) автоматическая генерация меток тем. Далее в нашей статье мы более подробно обсудим первые два этапа.

Четвертый этап включал сбор статистической информации о корпусе, автоматический подсчет количества токенов в текстах отдельных сборников с помощью счетчика, реализованного на языке *Python*.

Таким образом, процедура составления корпуса ТКиКЛ является трудоемким процессом, который включает в себя несколько этапов, начиная от сбора и лемматизации текстов до их систематизации. Благодаря использованию программных инструментов, этот процесс был частично автоматизирован и упрощен.

Структура корпуса включает три каталога: *Corpora*, *IMS*, *full_corpus*. Первые два каталога подразделяются на еще два с необработанными и лемматизированными текстами: *Corpora_raw*, *Corpora_lemmatized*, *IMS_raw*, *IMS_lemmatized*. Каждый из этих каталогов включает папки годов с файлами статей соответствующих сборников. Названия папок маркированы тегами по следующему шаблону: для неразмеченных текстов – *year*, *year_thesis* (например, *2004*, *2004_thesis*); для лемматизированных – *year_lem*, *year_thesis_lem* (*2004_lem*, *2004_thesis_lem*). Сами файлы унифицированы по шаблону: для неразмеченных текстов – *surname_conference name_year* / *thesis_year.txt* (например, *Gerd_CL_2006.txt*, *Gerd_CL_thesis_2004.txt*, *Masevich_IMS_2018.txt*), для лемматизированных – *surname_conference name_year* / *thesis year_lem.txt* (*Gerd_CL_2011_lem.txt*, *Alexeeva_IMS_2015_lem.txt*).

Каталог с тегом *IMS* включают в себя 11 папок (с маркировкой от *2013* до *2023* г.), с тегом *Corpora* – 12 папок (*2002*, *2004*, *2004_thesis*, *2005*, *2006*, *2008*, *2011*, *2013*, *2015*, *2017*, *2019*, *2021*). Сегмент корпуса, представляющий материалы конференции *Corpora*, в общей совокупности составили 442 файлов, материалы семинара *IMS* по компьютерной лингвистике и вычислительным онтологиям — 201 файл. Общий размер корпуса — более 1 млн. токенов.

Каталог *full_corpus* содержит 643 файла – все лемматизированные тексты двух конференций. Более подробное описание корпуса можно видеть в таблице 1, где указаны сборники – источники материалов, а также их количественный состав.

Таблица 1. Количественный состав корпуса ТКиКЛ

№	Конференция – год	Число текстов	Количество токенов
1	Корпусная лингвистика и лингвистические базы – 2002	21	62634
2	Корпусная лингвистика – 2004 / тезисы	26 / 44	65412 / 15237
3	MegaLing – 2005	18	27277
4	Корпусная лингвистика – 2006	40	55524
5	Корпусная лингвистика – 2008	40	57388
6	Корпусная лингвистика – 2011	48	47749
7	Корпусная лингвистика – 2013	36	45484
8	Корпусная лингвистика – 2015	42	46963
9	Корпусная лингвистика – 2017	49	43517
10	Корпусная лингвистика – 2019	45	58828
11	Корпусная лингвистика – 2021	33	47835
		442	541378
12	IMS CompLing – 2013	48	111284
13	IMS CompLing – 2014	54	133133
14	IMS CompLing – 2015	12	30692
15	IMS CompLing – 2016	7	15323
16	IMS CompLing – 2017	19	43217
17	IMS CompLing – 2018	16	37396
18	IMS CompLing – 2019	14	41315
19	IMS CompLing – 2020	8	18552
20	IMS CompLing – 2021	7	16167
21	IMS CompLing – 2022	7	16452
22	IMS CompLing – 2023	9	22385
		201	485916

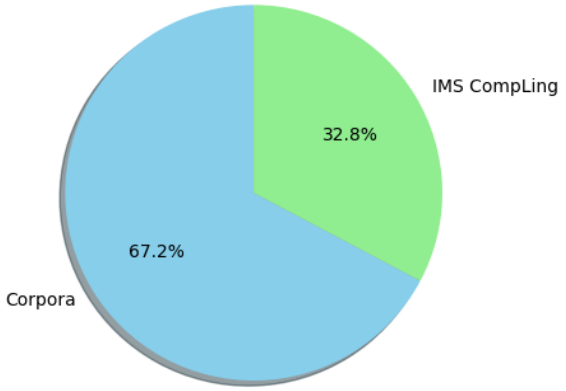

Рис. 1. Соотношение числа текстов в двух сегментах корпуса *Corpora* и *IMS CompLing*

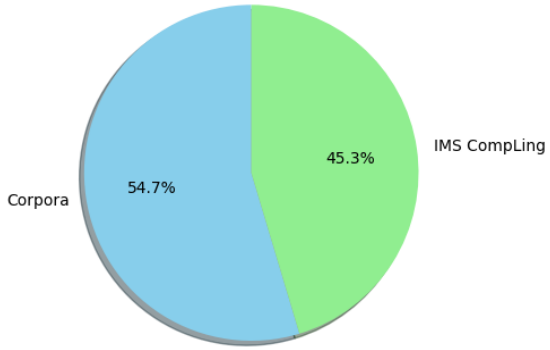

Рис. 2. Соотношение количества токенов в двух сегментах корпуса *Corpora* и *IMS CompLing*

На основе данных из таблицы 1 и рисунков 1-2 можно сделать следующие выводы.

1. Из круговой диаграммы сравнения числа текстов (рис. 1) видно, что объем сегмента Corpora более чем в 2 раза выше, чем соответствующий сегмент *IMS CompLing* (статей *IMS CompLing* меньше, чем статей *Corpora*).
2. По круговой диаграмме сравнения количества токенов (рис. 2) можно сделать вывод, что тексты сегмента *IMS CompLing* содержат больше токенов, чем тексты сегмента *Corpora* (статьи *IMS CompLing* длиннее статей *Corpora*).
3. Согласно распределению текстов в корпусе по годам, наибольшее количество текстов было опубликовано в 2017 году (49 текстов), наименьшее количество текстов было опубликовано в 2005 году (18 текстов); общее количество текстов в корпусе имеет тенденцию к увеличению с течением времени.
4. Согласно распределению количества токенов в корпусе по годам, наибольшее количество токенов было внесено в корпус по текстам 2004 года (65412 токенов), наименьшее количество токенов было внесено в корпус по текстам 2015 года (46963 токенов); общее количество токенов в корпусе также увеличивается со временем. Дополнительно можно отметить, что в 2013 году вклад материалов конференции *IMS CompLing* в корпус был наибольшим.
5. Наибольшее среднее количество токенов (средний объем текстов статей в токенах) было зарегистрировано в 2004 году (3101.88), а наименьшее в 2011 году (994.77). В целом, наблюдается некоторая вариативность в значениях за исследуемый период, однако общая тенденция к изменению колеблется в пределах от 1388.10 до 3101.88.

В ходе составления корпуса были выявлены некоторые проблемы, связанные с его структурой и сбором материалов. Одной из основных проблем является необходимость восстановления отсутствующих компонентов текста: шаблоны оформления текстов статей менялись, отдельные статьи не имеют авторской аннотации и не содержат авторские наборы ключевых слов и словосочетаний. Для решения данной проблемы была проведена генерация ключевых выражений и аннотаций с применением моделей машинного обучения (см. разделы 3 и 4 данной статьи). Еще одной проблемой является отсутствие сопоставимого подкорпуса со статьями на английском языке, поскольку англоязычные тексты составляют значительную часть трудов конференций *Corpora* и *IMS CompLing*. Создание такого подкорпуса является задачей следующего этапа работы с корпусом, направленного на улучшение качества и репрезентативности корпуса для проведения дальнейших исследований.

III. Автоматическая генерация ключевых слов и словосочетаний в текстах корпуса ТКиКЛ

Автоматическое выделение ключевых слов и словосочетаний является необходимой процедурой в процессе подготовки научного текста, способствующей формированию информационно-поискового портрета текста. Наборы ключевых выражений помогают быстро оценить содержание текстов в ходе индексирования, рубрикации, суммаризации, упрощения и перифразирования [5-10]. Методы выделения ключевых выражений разрабатывались прежде всего применительно к научным текстам с высокой концентрацией терминов и терминосочетаний, это объясняет ориентированность процедур автоматического выделения ключевых выражений на использование статистических признаков ключевых выражений, их структурной и лексико-грамматической организации (униграммы, биграммы, триграммы и т.д.), способы их ранжирования (регистрация их локализации в тексте, длина, встречаемость в составе других n-грамм), наличие одного корпуса текстов или пары корпусов – основного и фонового, возможность использования размеченных данных для организации процедур машинного обучения и т.д.

Автоматизация выделения ключевых выражений, равно как и ручная их разметка, является предметом дискуссий. Возникающие вопросы связаны с возможным несоответствием лексических единиц в реферативной и основной частях документа: зачастую назначаемые авторами ключевые выражения редко встречаются в тексте или вовсе в нем отсутствуют. В таких случаях неизбежно применение автоматических методов обработки данных. Базовыми количественными

характеристиками, по которым можно оценить потенциальную значимость ключевых выражений для читателя, являются их плотность (отношение частоты употребления в тексте по отношению к его общему объему) и пространственно-позиционные признаки (расположение в документе). Принято считать, что наиболее информативны выражения, встречающиеся в заголовке, аннотации, в начальной части текста (первый абзац, первые несколько предложений), а также в конце текста (в заключении) [6].

В сборниках *CL2002–2008*, согласно использованному издательскому шаблону, отсутствовали авторские ключевые выражения, это обуславливает необходимость восстановления существующих лакун для унификации представления текстов в корпусе ТКиКЛ. Для автоматического извлечения ключевых выражений в нашем исследовании рассматривались разнородные алгоритмы, а именно, статистические: Log-Likelihood, TF-IDF, Хи-квадрат; гибридные (лингвостатистические): RAKE, YAKE, MultiRAKE, PullEnti, RuTermExtract, графовые: TopicRank, с использованием машинного обучения: Spacy, KeyBERT [7-10]. Рассматриваемый набор методов не является исчерпывающим. При отборе методов выделения ключевых выражений мы учитывали возможность их применения в работе с русскоязычными текстами, в также возможность извлечения n-грамм разной структуры (униграмм, биграмм, триграмм и т.д.). Основные методы выделения ключевых выражений учитывают не только их типичность для определенного документа или классов документов, но и их коллокационную природу, что важно для терминологически насыщенных текстов.

В [7] приведен анализ наборов ключевых выражений, выделенных в 50 текстах корпуса ТКиКЛ с учетом пространственно-позиционных и стилистически детерминированных характеристик ключевых выражений. В результате серии экспериментов были сопоставлены эталонные ключевые выражения, выделенные экспертами из первого сегмента текстов, и ключевые выражения, извлеченные из второго сегмента автоматическими методами. Наилучшие результаты показали алгоритмы PullEnti, RAKE и RuTermExtract. В [10] было проведено сравнение алгоритмов генерации ключевых выражений для аннотаций научных статей, в ходе которого было установлено, что самые высокие результаты по F-мере показывают алгоритмы YAKE и TopicRank. Полученные данные были применены в проекте НейроКРЯ по разметке ключевых выражений в корпусе региональной прессы, где был применен алгоритм RuTermExtract, хорошо зарекомендовавший себя в предыдущих экспериментах. Следуя нашим наблюдениям и опыту НейроКРЯ, мы приняли решение о генерации ключевых выражений для статей в сборниках *CL2002–2008* с помощью алгоритма RuTermExtract, отбирая первые три слова и три словосочетания с наибольшим весом. Пример разметки приведен в таблице 2.

Таблица 2. Примеры разметки ключевых выражений в текстах корпуса ТКиКЛ

№	Статья	Ключевые слова	Ключевые словосочетания
1	Апресян Ю.Д., Иомдин Л.Л., Санников А.В., Сизов В.Г., Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004.	слово дескриптор корпус	семантическая информация семантический словарь семантическая роль
2	Захаров В.П., Толбаст С.П. Поисковая система сети Интернет и корпусные исследования // MegaLing 2005: Прикладная лингвистика в поисках новых путей. СПб., 2005.	интернет поиск запрос	поисковая система русский язык корпусные исследования
3	Зубов А.В. Корпус текстов белорусского языка // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006.	текст корпус кодирование	белорусский язык корпус текстов письменный текст
4	Герд А.С. Академическая лексикография как система корпусов // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006.	словарь слово значение	словарный грамматика академический словарь теоретическая семантика
5	Падучева Е.В. Прямая и косвенная диатеза ментального глагола: корпусное исследование // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008.	глагол мнение знание	рематический акцент прямая диатеза пропозициональный актант

IV. Автоматическая генерация аннотаций в текстах корпуса ТКиКЛ

Аннотация – это важный компонент структуры научной статьи, представляющий краткое и лаконичное изложение основного содержания исследования. Она представляет собой своего рода краткое резюме, которое помогает читателям быстро оценить, насколько статья соответствует их интересам и ожиданиям [11-13]. Структура аннотации, как правило, должна соответствовать требованиям IMRAD (Introduction, Methods, Results, and Discussion). Качественная аннотация научной статьи должна содержать следующие элементы: краткое изложение цели исследования, а также основных вопросов или задач, решаемых в работе; упоминание основных методов исследования, используемых для достижения поставленной цели; краткое описание основных результатов исследования или выводов, которые были получены в ходе работы. Помимо этого, аннотация может содержать уточнение, почему проведенное исследование важно, какие у него дальнейшие перспективы и какие практические или теоретические выводы можно сделать на его основе. Таким образом, аннотация должна быть ограниченного объема (в случае настоящего проекта до 250 слов) и информативной с точки зрения передачи содержания исходного текста, при этом структурированной и написанной доступным языком, чтобы читатели могли быстро понять суть исследования, не читая всю статью.

Автоматическая суммаризация текста широко применяется в различных областях, таких как информационные технологии, медицина, финансы, новости и другие, где требуется обработка большого объема информации для получения краткого обзора или анализа. Этот процесс может осуществляться с использованием различных методов и алгоритмов, предполагая суммаризация по предложениям, по документам, по корпусу текстов, по аспектам, одноязычную или многоязычную суммаризацию. Суммаризация представляет собой вариант семантической компрессии, в ходе которой исходное содержание передается в тексте с сокращением плана выражения, при этом сходные механизмы используются при упрощении, когда результирующий текст должен быть формально проще и не обязательно короче [14], и перифразировании, когда исходный и итоговый тексты должны характеризоваться сходными содержанием и формой [15]. Два основных подхода к созданию аннотации текста: экстрактивная и абстрактивная суммаризация [16]. Основное отличие между этими двумя подходами заключается в том, как они обрабатывают исходный материал и формируют краткое содержание текста. Экстрактивный подход к суммаризации предполагает извлечение наиболее важных фрагментов оригинального текста (предложений или фраз) и комбинирование этих фрагментов для построения аннотации. В основном, при использовании такого подхода сохраняется структура и форма оригинального текста, так как экстрактивная суммаризация не предполагает генерации новых текстов или перифразирования уже имеющихся текстов. В отличие от экстрактивной суммаризации, абстрактивный метод суммаризации не ограничивается извлечением предложений из исходного текста, а предполагает порождение нового текста ограниченного объема с заданным в оригинале содержанием. Абстрактивная суммаризация позволяет не только переформулировать исходные предложения, но и генерировать новые, которых нет в оригинальном тексте. Такой подход сложнее с точки зрения технологической реализации, так как требует понимания текста и способности на его основании генерировать новое содержание.

В корпусе ТКиКЛ отсутствовали аннотации для статей *CL 2002-2011*, по этой причине было принято решение сгенерировать их автоматически при помощи алгоритмов суммаризации. В настоящем проекте к задаче автоматической генерации аннотаций текстов научных статей были применены два алгоритма суммаризации: в качестве алгоритма экстрактивной суммаризации был выбран алгоритм, представленный в библиотеке *sumy* [17], абстрактивная суммаризация осуществлялась при помощи модели T5 семейства Трансформер *rut5_base_sum_gazeta* [18]. Аннотации различаются степенью подробности и объемом: как видно из Таблицы 3, аннотации *sumy* длиннее и конкретнее, тогда как аннотации *rut5_base_sum_gazeta* более краткие и обобщенные.

Таблица 3. Примеры аннотаций, сгенерированных для статей в текстах корпуса ТКиКЛ

	Статья	**Аннотация sumy**	**Аннотация rut5_base_sum_gazeta**
1	Ягунова Е.В. Исследование контекстной предсказуемости единиц текста с помощью корпусных ресурсов // Труды международной конференции «Корпусная лингвистика–2008». СПб, 2008. С. 396-403.	*В большей степени нас будут интересовать процедуры контекстной предсказуемости в рамках восприятия текста (речи), в меньшей степени мы обращаемся к данным порождения текста. Однако понятие «синтагматический сосед» требует уточнения; прежде всего, с точки зрения того, какая единица – словоформа или лемма –*	*С помощью корпусных ресурсов можно рассматривать механизмы контекстной предсказуемости в рамках восприятия текста. Это может быть связано с вероятностями влияния разных позиций, которые способствуют (или не способствуют) адекватному восприятию соответствующих единиц текста.*

		рассматривается в качестве коллоката. Таким образом, при решении разных задач контекстной предсказуемости оказывается важным сопоставлять данные по сочетаниям как словоформ, так и лексем.	

IV. Автоматическая разметка именованных сущностей в текстах корпуса ТКиКЛ

Именованные сущности (Named Entities – NE) — это слова или словосочетания, которые выделяют предметы или явления в ряде аналогичных предметов или явлений. Именованные сущности в текстах представляют собой конкретные организации, объекты, даты, места, и другие имена, которые имеют определенное значение и могут быть идентифицированы как отдельные субъекты. Задача распознавания именованных сущностей (Named Entity Recognition – NER) является важным этапом в извлечении информации (Information extraction, IE), которая состоит в автоматическом выделении структурированных данных из источников неструктурированной или слабоструктурированной информации и связана с информационным поиском и обработкой информации на естественных языках [19-22].

Существуют различные методы выделения именованных сущностей в текстах, например: с применением создаваемых вручную наборов правил, с применением специализированных парсеров (например, библиотека Natasha [23], Yargy-parser [24]), основанных на статистических моделях с применением классического и глубинного машинного обучения (например, модели NER в проекте DeepPavlov [25]).

Особые именованные сущности в текстах могут быть связаны с уникальными или специфическими объектами, событиями или понятиями, которые имеют особое значение или статус. Они играют важную роль в анализе текстов, так как они содержат ценную информацию о контексте и содержании текста. Их распознавание и классификация может помочь выделить ключевые аспекты текста и информацию определенного вида (например, термины, названия организаций, персоналии и т.д.) Стоит также учитывать, что набор именованных сущностей и связи между ними на уровне вложенных сущностей (Nested Entities) [26] будет существенно различаться в зависимости от типа текста и его тематики.

Существующие программные комплексы, библиотеки, программы и программные интерфейсы приложений (API) для решения задачи извлечения именованных сущностей охватывают широкий тематический диапазон текстов и предлагают общий, стандартный набор именованных сущностей. Например, программа Stanford NER (CRFClassifier) выделяет следующие типы NE: Person; Location; Organization; Date; Time; Money; Percentage [27].

Однако для решения задачи информационного поиска в тематических текстах необходим более развернутый набор тегов для выделения именованных сущностей с более подробной классификацией. В ходе анализа ключевых слов из текстов, вошедших в корпус ТКиКЛ, были выделены следующие особые именованные сущности, характерные для рассматриваемой предметной области. В Таблице 5 для каждого вида приведено название категории, возможный тег и примеры из корпуса. В ходе экспериментов с применением библиотек Natasha и yargy-парсера проведена разметка уникальных именованных сущностей в текстах корпуса ТКиКЛ.

Таблица 5. Уникальные именованные сущности в корпусе ТКиКЛ

	Тип уникальной именованной сущности	**Тег**	**Примеры**
1.	Названия конференций	CONF	*Диалог*
2.	Язык	LAN	*китайский язык, русский язык*
3.	Названия моделей	MODEL	*BERT, RuBERT*
4.	Проекты	PROJECT	*Текстометр, Русский конструктикон, RuSkell, CoCoCo*
5.	Стандарты	STANDARD	*CTB CNS, PKU*
6.	Форматы разметки	FORMAT	*CoNLL-U*
7.	Языки программирования	PR_LAN	*Python*
8.	Библиотеки	LIB	*UDPipe, Stanza*
9.	Алгоритмы	ALG	*fastHan, LTP, PKUSeg, Ckiptagger*
10.	Корпусы	CORP	*НКРЯ, подкорпус RU-AC, CyberCAT, ruTenTen11*
11.	Тесты	TEST	*Flesch Reading Ease, Flesch-Kincaid Grade*

VI. Тематическая рубрикация текстов в корпусе ТКиКЛ

В процессе формирования корпуса ТКиКЛ проводилась экспертная разметка текстов по рубрикам. Для этой цели была разработана схема рубрикации, содержащая темы работы секций конференций. Темы соответствуют

названиям, предложенными членами организационных комитетов конференций. Перед проведением экспертной разметки была осуществлена нормализация названий тем, результат представлен в Таблице 6. Данная экспертная тематическая разметка будет использована для верификации результатов кластеризации текстов и тематических моделей, обученных на корпусе.

Таблица 6. Темы в схеме рубрикации текстов корпуса ТКиКЛ

№	Схема рубрикации
1.	Общие вопросы корпусной лингвистики
2.	Создание, разработка и применения корпусов
3.	Статистические исследования на материале корпусов
4.	Корпусы и лексикография
5.	Морфология и синтаксис в корпусах
6.	Семантика в корпусах
7.	Обучающие корпусы
8.	Исторические корпусы
9.	Параллельные корпусы и машинный перевод
10.	Речевые и мультимедийные корпусы
11.	Корпусы художественных текстов

V. Результаты тематического моделирования корпуса ТКиКЛ

A. Методология тематического моделирования

Под тематическим моделированием традиционно понимается особый способ построения структурно-семантической модели корпуса текстов, которая определяет взаимосвязи тем, документов и слов-тематизаторов [28]. 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 [29], 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 [30], перплексии [31] и энтропии [32]. 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 [33–35]. 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 [36] 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.

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

Алгоритм NMF (Non-Negative Matrix Factorization, неотрицательная матричная факторизация) [37, 38] 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 [39]. 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.

C. 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 [40–42]. 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 [39]. 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.

D. 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 [43]. Как известно, тематическая модель LDA частично решает проблему переобучения pLSA [44].

В экспериментах использовалась реализация алгоритма LDA в библиотеках scikit-learn [39] и gensim [45]. 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*...).

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

Модель битермов (Biterm Topic Model) [46] 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 [47]. 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.

VI. Результаты генерации меток тем в корпусе ТКиКЛ

A. 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-транформеров, 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 [17, 20], 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 [48] 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, составленный 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-грамм по весу.

Постобработка включает фильтрацию n-грамм по следующим правилам: удаление повторов и совхождений; 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 7.

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. Кроме того, результаты не детерминированы, как и в случае нейросетей, и зависят от результатов работы поисковых систем, которые периодически обновляются. Разумеется, результат зависит также и от выбранной методики тематического

моделирования: в нашем исследовании лучший результат был получен для моделей LSA и NMF, где метки оказались более интепретируемы. Наконец, 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.

Таблица 7. 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*
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*
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*

B. 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 [34–36], мы применяли статические дистрибутивно-семантические модели типа Word2Vec [49] 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 8. 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 8. 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*
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*
модель, алгоритм, тематический, слово, метод, текст, документ, тема, оценка, слов, результат, анализ, задача, распределение,	*статистика, гипотеза, ранжирование, классификатор*	*lda, отзыв, классификатор, кластеризация*

количество, эксперимент, матрица, коллекция, вероятность, вероятностный…		

C. 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. 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 [36, 50]).

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, представленные в VI пункте выше. Рубрики из схемы экспертной рубрикации приведены в последней колонке Таблицы 9.

VII. Заключение

В результате подготовки нового корпусного ресурса, включающего материалы конференции «Корпусная лингвистика» с 2002 по 2021 гг., семинара «Компьютерная лингвистика и вычислительные онтологии» с 2011 по 2023 гг., а также некоторые другие материалы, были проведены следующие работы: преобразование текстов в формат *.txt, фильтрация нетекстовых элементов, разметка метаданных (авторы, аффилиации, названия, наборы ключевых выражений, аннотации, названия конференций, годы издания, тематические рубрики, именованные сущности и т.д.). Была проведена унификация формата описания структуры текста, восстановлены лакуны – сгенерированы наборы ключевых выражений с применением алгоритма RuTermExtract и аннотации с помощью алгоритмов экстрактивной и абстрактивной суммаризации.

Разработка специализированных корпусов текстов, к которым относится корпус ТКиКЛ, 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.

Таблица 9. Примеры меток тем, сгенерированных ChatGPT

Темы (фрагмент выдачи)	**Метки ChatGPT**	**Главное слово в теме, выбор ChatGPT**	**Рубрика**
морфологический, синтаксический, разметка, слово, грамматический, форма, предложение, словоформа, разбор, существитель-ное, неоднозначность, автоматический, ошибка, парсер, омонимия, часть, анализатор, снятие, падеж, вариант	*"Морфологи-ческая и синтаксичес-кая разметка", "Граммати-ческий анализ предложений", "Автомати-ческий разбор текста и ошибки"*	*разметка*	*Морфологи я и синтаксис в корпусах*
учебный, студент, преподаватель, обучение, образовательный, курс, обучения, корпусный, английский, ошибка, студентов, задание, учиться, технология,	*"Обучение и образование с использовани-ем корпусов", "Английский язык в образователь-ной среде", "Корпусный анализ ошибок студентов"*	*учебный*	*Обучаю-щие корпусы*

материал, использование, тест, профессиональный, возможность, работа			

Библиография

[1] О.А. Митрофанова, В.П. Захаров, Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 321 – 328. URL: https://www.dialog-21.ru/digests/dialog2009/materials/pdf/49.pdf (дата обращения: 25.11.2024).

[2] Н.В. Виноградова, О.А. Митрофанова, Формальная онтология как инструмент систематизации данных в русскоязычном корпусе текстов по корпусной лингвистике // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008. URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (дата обращения: 25.11.2024).

[3] Н.В. Виноградова, О.А. Митрофанова, П.В. Паничева, Автоматическая классификация терминов в русскоязычном корпусе текстов по корпусной лингвистике // Труды девятой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции» (RCDL–2007). Переславль-Залесский: 2007. URL : http://rcdl.ru/doc/2007/paper_31_v1.pdf (дата обращения: 25.11.2024).

[4] В.П. Захаров, С.Ю. Богданова, Корпусная лингвистика. СПб., 2020.

[5] Е.В. Тихонова, М.А. Косычева, Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology, 2022. Вып. 3 (4). P. 7 – 15. URL: https://elibrary.ru/item.asp?id=49446588 (дата обращения: 25.11.2024).

[6] O. Kamshilova, L. Beliaeva, L. Geikhman, Author's Choice for Keyword List: Research Aspect // PRLEAL-2019. R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL–2019). CEUR Workshop Proceedings. Saint Petersburg, Russia, November 27, 2019. 2020. P. 47–59. URL: https://elibrary.ru/item.asp?id=42584043 (дата обращения: 25.11.2024).

[7] О.А. Митрофанова, Д.А. Гаврилик, Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Вып. 13 (4). С. 22 – 40. URL: https://elib.spbstu.ru/dl/2/j23-158.pdf/en/info (дата обращения: 25.11.2024).

[8] Д.Д. Гусева, О.А. Митрофанова, Ключевые выражения в русскоязычных научно-популярных текстах: сравнение восприятия устной и письменной речи с результатами автоматического анализа // Terra Linguistica. 2024 Вып. 15 (1). С. 20 – 35.

[9] A. Moskvina, E. Sokolova, O. Mitrofanova, 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 / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS. P. 369–372. URL: https://elibrary.ru/item.asp?id=41112843 (дата обращения: 25.11.2024).

[10] Д.А. Морозов и др., Генерация ключевых слов для аннотаций русскоязычных научных статей / Морозов Д.А., Глазкова А.В., Тютюльников М.А., Иомдин Б.Л. // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2023. №1.

[11] A. Aries, D. Zegour, H. Walid, Automatic Text Summarization: What has been done and what has to be done // arXiv:1904.00688. 2019. P. 1 – 34. URL: https://arxiv.org/abs/1904.00688 (дата обращения: 25.11.2024).

[12] A. Nenkova, K. McKeown, Automatic Summarization // Foundations and Trends in Information Retrieval. 2011. Vol. 5 (2-3). P. 103 – 233. Available: https://core.ac.uk/download/pdf/76383212.pdf (дата обращения: 25.11.2024).

[13] M. Allahyari et al., Text Summarization Techniques: a Brief Survey / Allahyari M., Pouriyeh S., ssefi M., Safaei S., Trippe E.D., Gutierrez J.B., Kochut K. // arXiv preprint arXiv:1707.02268. 2017. URL: https://arxiv.org/abs/1707.02268 (дата обращения: 25.11.2024).

[14] M. Athugodage, O. Mitrofanova, V. Gudkov, Transfer Learning for Russian Legal Text Simplification // Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024. 2024. P. 59 – 69. URL: https://aclanthology.org/2024.readi-1.6/ (дата обращения: 25.11.2024).

[15] V. Gudkov, O. Mitrofanova, E. Filippskikh, Automatically Ranked Russian Paraphrase Corpus for Text Generation // Proceedings of the Fourth Workshop on Neural Generation and Translation. Association for Computational Linguistics, 2020. P. 54 – 59. URL: https://aclanthology.org/2020.ngt-1.6/ (дата обращения: 25.11.2024).

[16] J. Pilault et al. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models / Pilault J., Li R., Subramanian S., Pal C. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P. 9308 – 9319. URL: https://aclanthology.org/2020.emnlp-main.748/ (дата обращения: 25.11.2024).

[17] Automatic text summarizer. URL: https://pypi.org/project/sumy/ (дата обращения: 25.11.2024).

[18] RuT5SumGazeta. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (дата обращения: 25.11.2024

[19] M.M. Tikhomirov, N.V. Loukachevitch, B.V. Dobrov, Recognizing Named Entities in Specific Domain // Lobachevskii Journal of Mathematics. Vol. 41 (8). 2020. P. 1591 – 1602. URL: https://link.springer.com/article/10.1134/S199508022008020X (дата обращения: 25.11.2024).

[20] Д.М. Костюк, Н.К. Широков, Методы идентификации именованных сущностей в задачах обработки потока научных новостей // Менеджмент вузовских библиотек. Минск, 2021. С. 50 – 54. URL: https://elibrary.ru/item.asp?id=49171334 (дата обращения: 25.11.2024).

[21] А.А. Навроцкий, Е.В. Кривальцевич, Сравнительный анализ систем извлечения именованных сущностей из неструктурированных публицистических текстов // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня. Минск, 2020. С. 12 – 18. URL: https://elibrary.ru/item.asp?id=43934323 (дата обращения: 25.11.2024).

[22] V. Yadav, S. Bethard, A Survey on Recent Advances in Named Entity Recognition from Deep Learning models // Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 2018. P. 2145 – 2158. URL: https://arxiv.org/abs/1910.11470 (дата обращения: 25.11.2024).

[23] Natasha // GitHub Repository. URL: https://github.com/natasha/natasha (дата обращения: 26.03.2024).

[24] Yargy // GitHub Repository. URL: https://github.com/natasha/yargy (дата обращения: 25.11.2024).

[25] Named Entity Recognition (NER) // DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (дата обращения: 25.11.2024).

[26] NEREL // GitHub Repository. URL: https://github.com/nerel-ds/NEREL (дата обращения: 25.11.2024).

[27] Stanford NER. URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (дата обращения: 25.11.2024).

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

[29] A. Moskvina, E. Sokolova, O. Mitrofanova, 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. P. 369–372.

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

[31] G. Heinrich, Parameter Estimation for Text Analysis // Technical Report. 2005. P. 1–32.

[32] S. Koltcov, Application of Rényi and Tsallis Entropies to Topic Modeling Optimization // Physica A: Statistical Mechanics and its Applications. 2018. 512. P. 1192–1204.
[33] A. Erofeeva, O. Mitrofanova, Automatic Assignment of Labels in Topic Modeling for Russian Corpora // Structural and Applied Linguistics. Volume 12. St. Petersburg, 2019. P. 122–147.
[34] A. Kriukova, A. Erofeeva, O. Mitrofanova, K. Sukharev, 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.
[35] O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov, 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, vol. 1357. Springer, 2021. P. 102−114.
[36] O.A. Mitrofanova, M.M. Athugodage, L.V. Ten, Topic Label Generation in the Popular Science Corpus // 26th international conference «Internet and Modern Society» (IMS–2023): International Workshop «Computational Linguistics» (CompLing 2023). Proceedings. Springer Nature. 2023.
[37] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina, 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.
[38] D. Kuang, J. Choo, H. Park, Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering // Partitional Clustering Algorithms. 2015. P. 215–243.
[39] Scikit-Learn. URL: https://scikit-learn.org/ (дата обращения: 25.11.2024).
[40] T.K. Landauer, P.W. Foltz, D. Laham, Introduction to Latent Semantic Analysis // Discourse Processes. Issue 25. 1998. P. 259–284.
[41] А.В. Чижик, 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.
[42] М.А. Кирина, 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. 20 (2). С. 93–109.
[43] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, University of California, Berkeley, Berkeley, CA 94720. 2002. P. 993–1022.
[44] T. Hofmann, Probabilistic Latent Semantic Indexing // ACM SIGIR Forum. Vol. 51:2. 2017. P. 211–218.
[45] Gensim. URL: https://radimrehurek.com/gensim/ (дата обращения: 25.11.2024).
[46] X. Yan, J. Guo, Y. Lan, X. Cheng, A Biterm Topic Model for Short Texts // WWW 2013. Proceedings of the 22nd International Conference on World Wide Web. 2013. P. 1445–1456.
[47] Biterm. URL: https://pypi.org/project/biterm/ (дата обращения: 25.11.2024).
[48] Google. URL: https://www.google.ru/ (дата обращения: 25.11.2024).
[49] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space // URL: https://arxiv.org/abs/1301.3781 (дата обращения: 25.11.2024).
[50] О.А. Митрофанова, 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., Санкт-Петербург. СПб., 2024.

Митрофанова Ольга Александровна, Санкт-Петербургский государственный университет, канд. филол. наук, доцент, ORCID 0000-0002-3008-5514 (e-mail: o.mitrofanova@spbu.ru)
Адамова Мария Антоновна, Санкт-Петербургский государственный университет (e-mail: st110061@student.spbu.ru)
Букреева Людмила Александровна, Санкт-Петербургский государственный университет (e-mail: st110502@student.spbu.ru)
Голубев Ростислав Васильевич, Санкт-Петербургский государственный университет (e-mail: st110682@student.spbu.ru)
Гусяцкая Полина Андреевна, Санкт-Петербургский государственный университет (e-mail: st068584@student.spbu.ru)
Зернова Алиса Кирилловна, Санкт-Петербургский государственный университет (e-mail: st068103@student.spbu.ru)
Литвинова Анна Артемовна, Санкт-Петербургский государственный университет (e-mail: st110228@student.spbu.ru)
Макеев Кирилл Владимирович, Санкт-Петербургский государственный университет (e-mail: st110200@student.spbu.ru)
Павликова Владислава Станиславовна, Санкт-Петербургский государственный университет (e-mail: st109999@student.spbu.ru)
Плюснина Елизавета Алексеевна, Санкт-Петербургский государственный университет (e-mail: st109958@student.spbu.ru)
Сологуб Полина Юрьевна, Санкт-Петербургский государственный университет (e-mail: st095317@student.spbu.ru)
Сухан Даниил Дмитриевич, Санкт-Петербургский государственный университет (e-mail: st110829@student.spbu.ru)
Трошина Александра Валерьевна, Санкт-Петербургский государственный университет (e-mail: st110338@student.spbu.ru)
Уткина Александра Алексеевна, Санкт-Петербургский государственный университет (e-mail: st110578@student.spbu.ru)

Data Mining in the Text Corpus on Corpus and Computational Linguistics

O.A. Mitrofanova, M.A. Adamova, L.A. Bukreeva, R.V. Golubev, P.A. Gusyatskaya, A.K. Zernova, K.V. Makeev, A.A. Litvinova, V.S. Pavlikova, E.P. Plyusnina, P.Ju. Sologub, D.D. Sukhan, A.V. Troshina, A.A. Utkina

Abstract— The article is dedicated to the challenges of creating a corpus of articles on corpus and computational linguistics, which is being developed at the Department of Mathematical Linguistics of St. Petersburg State University (SPBU). The corpus is compiled under the supervision of V.P. Zakharov and includes texts from the "Corpus Linguistics" conference reports from 2002 to 2021, the "Computational Linguistics and Computational Ontologies" seminar from 2011 to 2023, as well as some other materials. During the development of the corpus resource, standardization of text presentation format was carried out, and the structure of the articles was investigated. Experiments were carried out on the generation of keywords and annotations in cases where the original text did not contain this information. Types of named entities recorded in the corpus were examined, and an algorithm for their annotation was implemented. Analysis of distribution of conference reports between thematic blocks of the conferences was fulfilled according to the expert annotation scheme. The results of experiments on training a family of topic models (NMF, LSA, LDA, Biterm) on the text corpus are presented in the paper. Generalization 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 modeling with the assignment of topic labels are compared with data on the distribution of reports by conference thematic blocks in accordance with the expert markup scheme.

***Keywords*— corpus linguistics, conference materials, annotation, keywords, summaries, thematic annotation, named entities, topic modeling, topic labels, rubrication.**

REFERENCES

[1] O. A. Mitrofanova, and V. P. Zakharov, “Automatic Analysis of Terminology in the Russian Text Corpus on Corpus Linguistics,” in *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue 2009"* (Bekasovo, May 27-31, 2009), issue. 8(15), Moscow, RSUH, pp. 321 – 328, 2009, URL: https://www.dialog-21.ru/digests/dialog2009/materials/pdf/49.pdf (accessed date: 25.11.2024).

[2] N. V. Vinogradova, and O. A. Mitrofanova, “Formal Ontology as a Tool for Systematizing Data in the Russian Text Corpus on Corpus Linguistics,” in *Proceedings of the International Conference "Corpus Linguistics - 2008"*, St. Petersburg, 2008, URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (date of access: 25.11.2024).

[3] N. V. Vinogradova, O. A. Mitrofanova, and P. V. Panicheva, “Automatic Classification of Terms in the Russian Text Corpus on Corpus Linguistics,” in *Proceedings of the Ninth All-Russian Scientific Conference "Electronic Libraries: Advanced Methods and Technologies, Electronic Collections" (RCDL-2007)*, Pereslavl-Zalessky, 2007, URL: http://rcdl.ru/doc/2007/paper_31_v1.pdf (date of access: 25.11.2024).

[4] V. P. Zakharov, and S.Yu. Bogdanova, “Corpus Linguistics”, St. Petersburg, 2020.

[5] E. V. Tikhonova, and M. A. Kosycheva, “Effective Keyword(s): Formulation Strategies,” *Health, Food & Biotechnology*, issue 3(4), pp. 7–15, 2022, URL: https://elibrary.ru/item.asp?id=49446588 (accessed date: 25.11.2024).

[6] O. Kamshilova, L. Beliaeva, and L. Geikhman, “Author’s Choice for Keyword List: Research Aspect,” in *R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL–2019)*, CEUR Workshop Proceedings, Saint Petersburg, Russia, November 27, 2019, pp. 47–59, 2020, URL: https://elibrary.ru/item.asp?id=42584043 (accessed date: 25.11.2024).

[7] O. A. Mitrofanova, and D. A. Gavrilik, “Experiments on Automatic Extraction of Key Expressions in Stylistically Diverse Corpora of Russian Text Corpora,” *Terra Linguistica*, issue 13(4), pp. 22–40, 2022, URL: https://elib.spbstu.ru/dl/2/j23-158.pdf/en/info (accessed date: 25.11.2024).

[8] D. D. Guseva, and O. A. Mitrofanova, “Key Expressions in Russian Popular Science Texts: Comparison of Oral and Written Speech Perception with the Results of Automatic Analysis,” Terra Linguistica, issue 15(1), pp. 20–35, 2024.

[9] A. Moskvina, E. Sokolova, and O. Mitrofanova, “KeyPhrase Extraction from the Russian Corpus on Linguistics by means of KEA and RAKE Algorithm,” in *Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL’2018*, October 9–12, 2018, Moscow, Russia, Conference Proceedings, ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, and V. Sukhomlin, FRC CSC RAS, pp. 369 – 372, 2018, URL: https://elibrary.ru/item.asp?id=41112843 (accessed date: 25.11.2024).

[10] D. A. Morozov, et al., “Generation of Keywords for Abstracts of Russian Scientific Articles,” Morozov D.A., Glazkova A.V., Tyutulnikov M.A., Iomdin B.L., *Bulletin of NSU. Series: Linguistics and intercultural communication*, no. 1, 2023.

[11] A. Aries, D. Zegour, and H. Walid, “Automatic Text Summarization: What has been done and what has to be done,” arXiv:1904.00688, pp. 1–34, 2019, URL: https://arxiv.org/abs/1904.00688 (accessed date: 25.11.2024).

[12] A. Nenkova, and K. McKeown, “Automatic Summarization,” *Foundations and Trends in Information Retrieval*, vol. 5(2-3), pp. 103–233, 2011, URL: https://core.ac.uk/download/pdf/76383212.pdf (accessed date: 25.11.2024).

[13] M. Allahyari, et al., “Text Summarization Techniques: a Brief Survey,” Allahyari M., Pouriyeh S., ssefi M., Safaei S., Trippe E.D., Gutierrez J.B., and Kochut K., *arXiv preprint*, 2017, URL: https://arxiv.org/abs/1707.02268 (accessed date: 25.11.2024).

[14] M. Athugodage, O. Mitrofanova, and V. Gudkov, “Transfer Learning for Russian Legal Text Simplification,” in *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pp. 59–69, 2024, URL: https://aclanthology.org/2024.readi-1.6/ (accessed date: 25.11.2024).

[15] V. Gudkov, O. Mitrofanova, and E. Filippskikh, “Automatically Ranked Russian Paraphrase Corpus for Text Generation,” in *Proceedings of the Fourth Workshop on Neural Generation and Translation. Association for Computational Linguistics*, pp. 54–59, 2020, URL: https://aclanthology.org/2020.ngt-1.6/ (accessed date: 25.11.2024).

[16] J. Pilault, et al., “On Extractive and Abstractive Neural Document Summarization with Transformer Language Models,” Pilault J., Li R., Subramanian S., and Pal C., in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 9308–9319, 2020,

URL: https://aclanthology.org/2020.emnlp-main.748/ (accessed date: 25.11.2024).
[17] *Automatic Text Summarizer*, URL: https://pypi.org/project/sumy/ (accessed date: 25.11.2024).
[18] *RuT5SumGazeta*, URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (accessed date: 25.11.2024
[19] M. M. Tikhomirov, N. V. Loukachevitch, and B. V. Dobrov, "Recognizing Named Entities in Specific Domain," *Lobachevskii Journal of Mathematics*, vol. 41(8), pp. 1591–1602, 2020, doi: 10.1134/S199508022008020X.
[20] D. M. Kostyuk, and N. K. Shirokov, "Methods for Identifying Named Entities in the Tasks of Processing the Flow of Scientific News," in *Management of University Libraries*, Minsk, pp. 50–54, 2021, URL: https://elibrary.ru/item.asp?id=49171334 (accessed date: 25.11.2024).
[21] A. A. Navrotsky, and E. V. Krivaltsevich, "Comparative Analysis of Systems for Extracting Named Entities from Unstructured Journalistic Texts," in *BIG DATA and Advanced Analytics = BIG DATA and high-level analysis*, Minsk, pp. 12–18, 2020, URL: https://elibrary.ru/item.asp?id=43934323 (accessed date: 25.11.2024).
[22] V. Yadav, and S. Bethard, "A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Association for Computational Linguistics, pp. 2145–2158, 2018, URL: https://arxiv.org/abs/1910.11470 (accessed date: 25.11.2024).
[23] Natasha, *GitHub Repository*, URL: https://github.com/natasha/natasha (accessed date: 02.02.2024).
[24] Yargy, *GitHub Repository*, URL: https://github.com/natasha/yargy (accessed date: 25.11.2024).
[25] Named Entity Recognition (NER), *DeepPavlov*, URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (accessed date: 25.11.2024).
[26] NEREL, *GitHub Repository*, URL: https://github.com/nerel-ds/NEREL (accessed date: 25.11.2024).
[27] *Stanford NER*, URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (accessed date: 25.11.2024).
[28] K. V. Vorontsov, "Probabilistic Topic Modeling: ARTM Regularization Theory and the BigARTM Open Source Library," *URSS*, 2023.
[29] A. Moskvina, E. Sokolova, and O. Mitrofanova, "KeyPhrase Extraction from the Russian Corpus on Linguistics by Means of KEA and RAKE Algorithm," in *Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018*, October 9–12, 2018, Moscow, Russia, FRC CSC RAS, pp. 369–372.
[30] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262−272, 2011.
[31] G. Heinrich, "Parameter Estimation for Text Analysis," *Technical Report*, pp. 1–32, 2005.
[32] S. Koltcov, "Application of Rényi and Tsallis Entropies to Topic Modeling Optimization," *Physica A: Statistical Mechanics and its Applications*, no. 512, pp. 1192–1204, 2018.
[33] A. Erofeeva, and O. Mitrofanova, "Automatic Assignment of Labels in Topic Modeling for Russian Corpora," *Structural and Applied Linguistics*, vol. 12, pp. 122–147, 2019.
[34] A. Kriukova, A. Erofeeva, O. Mitrofanova, and K. Sukharev, "Explicit Semantic Analysis as a Means for Topic Labeling," in *Artificial Intelligence and Natural Language Processing: 7th International Conference*, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings. Springer, Cham, pp. 167–177, 2018.
[35] O. Mitrofanova, A. Kriukova, V. Shulginov, and V. Shulginov, "E-hypertext Media Topic Model with Automatic Label Assignment," in *Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference*, AIST 2020, Revised Supplementary Proceeding, *Communications in Computer and Information Science*, vol. 1357, Springer, pp. 102−114, 2021.
[36] O. A. Mitrofanova, M. M. Athugodage, and L. V. Ten, "Topic Label Generation in the Popular Science Corpus," in *26th international conference «Internet and Modern Society» (IMS–2023), International Workshop «Computational Linguistics» (CompLing 2023)*, Proceedings, Springer Nature, 2023.
[37] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, and M. Kirina, "Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction," in *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence*, MICAI 2020, vol. 12469, pt. 2, P. 134–152, 2020.
[38] D. Kuang, J. Choo, and H. Park, "Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering," *Partitional clustering algorithms*, pp. 215–243, 2015.
[39] Scikit-Learn, URL: https://scikit-learn.org/ (accessed date: 25.11.2024).
[40] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to Latent Semantic Analysis," *Discourse Processes*, issue 25, pp. 259–284, 1998.
[41] A. V. Chizhik, "Using Topic Modeling Methods to Assess the Degree of Media Influence on Public Mood," in *Computational Linguistics and Computational Ontologies*, issue 5, Proceedings of the XXIV International United Scientific Conference "Internet and Modern Society", IMS-2021, St. Petersburg, June 24–26, 2021, SPb., ITMO University, pp. 70–78, 2021.
[42] M. A. Kirina, "Comparison of Topic Models Based on LDA, STM, and NMF for Qualitative Analysis of Russian Short Fiction," *Bulletin of the Novosibirsk State University. Series: Linguistics and Intercultural Communication*, no. 20(2), pp. 93–109, 2022.
[43] D. M. Blei, A.Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," University of California, Berkeley, Berkeley, CA 94720, pp. 993–1022, 2002.
[44] T. Hofmann, "Probabilistic Latent Semantic Indexing," *ACM SIGIR Forum*, vol. 51,2, pp. 211–218, 2017.
[45] *Gensim*, URL: https://radimrehurek.com/gensim/ (accessed date: 25.11.2024).
[46] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A Biterm Topic Model for Short Texts," in *WWW 2013. Proceedings of the 22nd International Conference on World Wide Web*, pp. 1445–1456, 2013.
[47] *Biterm*, URL: https://pypi.org/project/biterm/ (accessed date: 25.11.2024).
[48] *Google*, URL: https://www.google.ru/ (accessed date: 25.11.2024).
[49] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," URL: https://arxiv.org/abs/1301.3781 (accessed date: 25.11.2024).
[50] O. A. Mitrofanova, "Search and Ranking of Texts in a Special Corpus based on Topic Modeling," in *Proceedings of the International Conference "Corpus Linguistics - 2023"* (SPb Corpora 2023), june 21-23, 2023, St. Petersburg, SPb., 2024.

Mitrofanova Olga Aleksandrovna, St. Petersburg State University, Ph.D. in Philology, Associate Professor, ORCID 0000-0002-3008-5514 (e-mail: o.mitrofanova@spbu.ru)
Adamova Maria Antonovna, St. Petersburg State University (e-mail: st110061@student.spbu.ru)
Bukreeva Lyudmila Aleksandrovna, St. Petersburg State University (e-mail: st110502@student.spbu.ru)
Golubev Rostislav Vasilievich, St. Petersburg State University (e-mail: st110682@student.spbu.ru)
Gusyatskaya Polina Andreevna, St. Petersburg State University (e-mail: st068584@student.spbu.ru)
Zernova Alisa Kirillovna, St. Petersburg State University (e-mail: st068103@student.spbu.ru)
Litvinova Anna Artemovna, St. Petersburg State University (e-mail: st110228@student.spbu.ru)
Makeev Kirill Vladimirovich, St. Petersburg State University (e-mail: st110200@student.spbu.ru)
Pavlikova Vladislava Stanislavovna, St. Petersburg State University (e-mail: st109999@student.spbu.ru)
Plyusnina Elizaveta Alekseevna, St. Petersburg State University (e-mail: st109958@student.spbu.ru)
Sologub Polina Yuryevna, St. Petersburg State University (e-mail: st095317@student.spbu.ru)
Sukhan Daniil Dmitrievich, St. Petersburg State University (e-mail: st110829@student.spbu.ru)
Troshina Alexandra Valerievna, St. Petersburg State University (e-mail: st110338@student.spbu.ru)
Utkina Alexandra Alekseevna, St. Petersburg State University (e-mail: st110578@student.spbu.ru)