Корпус текстов по корпусной лингвистике: состав и этапы формирования

О. А. Митрофанова, М. А. Адамова, Л. А. Букреева, А. К. Зернова,
А. А. Литвинова, В. С. Павликова, П. Ю. Сологуб

Санкт-Петербургский государственный университет

o.mitrofanova@spbu.ru, st110061@student.spbu.ru,
st110502@student.spbu.ru, st068103@student.spbu.ru,
st110228@student.spbu.ru, st109999@student.spbu.ru,
st095317@student.spbu.ru

Аннотация

Статья посвящена проблемам разработки корпуса статей по корпусной 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 ходе работы над корпусным ресурсом 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.

Ключевые слова: корпусная лингвистика, материалы конференций, разметка, ключевые слова, аннотации, тематическая разметка, именованные сущности

Библиографическая ссылка: 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. 13–29. DOI: 10.17586/2541-9781-2024-8-13-29.

1. 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; 2; 3]. В данной статье рассматриваются следующие решенные нами задачи:

- формирование корпуса статей по корпусной лингвистике;
- типы информации, представленные в корпусе;
- разметка ключевых выражений в корпусе;
- генерация аннотаций статей;
- систематизация и разметка именованных сущностей.

Помимо этих задач были решены задачи систематизации рубрик в корпусе и мультимодальной тематической разметки, по формированию базы данных с метаинформацией и по разработке системы визуализации результатов поиска.

2. 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 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 процедуры:

- автоматическое выделение ключевых слов и выражений;
- автоматическая генерация аннотаций;
- тематическое моделирование;
- 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 файла, 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 и на рисунках 1–5, 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**

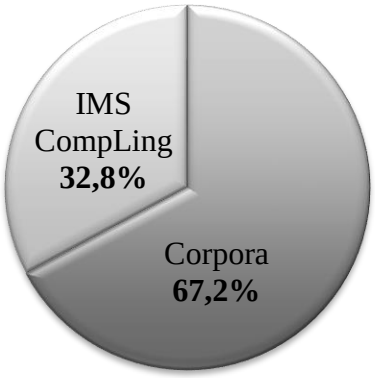

Рис. 1. Соотношение числа текстов в двух сегментах корпуса Corpora и IMS CompLing

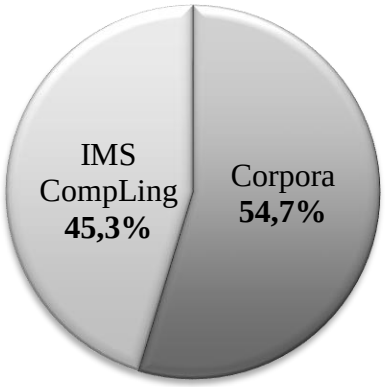

Рис. 2. Соотношение количества токенов в двух сегментах корпуса Corpora и IMS CompLing

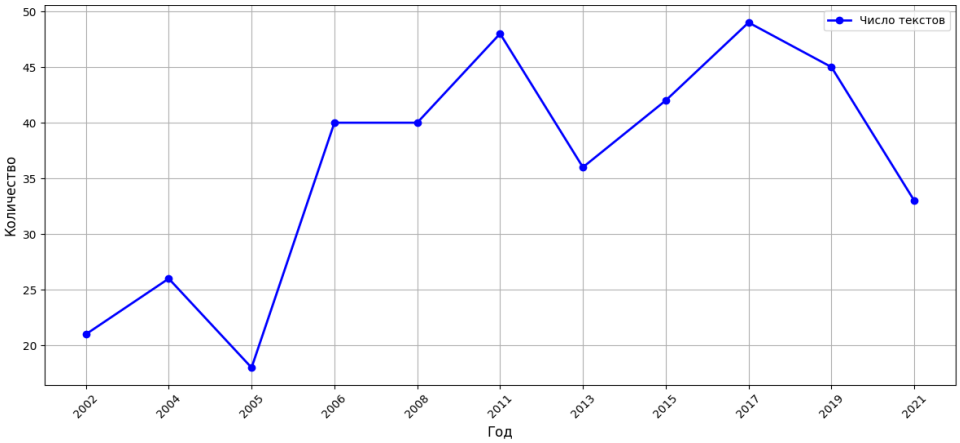

Рис. 3. Распределение текстов в корпусе по годам

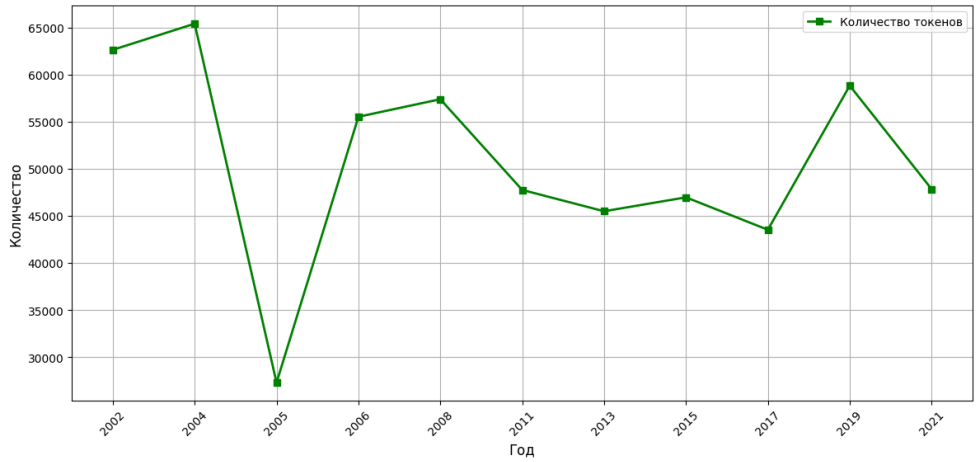

Рис. 4. Распределение токенов в корпусе по годам

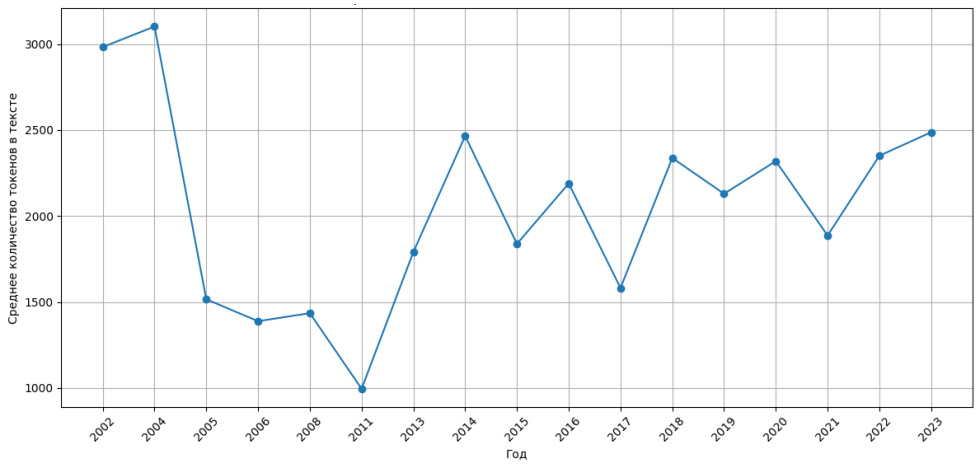

Рис. 5. Среднее количество токенов в тексте по годам

На основе данных из таблицы 1 и рисунков 1–5 можно сделать следующие выводы:

- 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);
- 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);
- согласно распределению текстов, в корпусе по годам (рис. 3), 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;
- согласно распределению количества токенов в корпусе по годам (рис. 4), 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);

- наибольшее среднее количество токенов (средний объем текстов статей в токенах) было зарегистрировано в 2004 году (3101.88), а наименьшее в 2011 году (994.77) (рис. 5); 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.

3. 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 перефразирования [5; 6; 7; 8; 9; 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; 8; 9; 10]. Рассматриваемый набор методов не является исчерпывающим. При отборе методов выделения ключевых выражений мы учитывали возможность их применения в работе с русскоязычными текстами, а 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	Андреев А. В. Архитектура информационно-поисковой системы для индоевропейского компьютерного тезауруса // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004	тезаурус, корпус, ИПС	индоевропейский компьютерный тезаурус, формат TEI, информационно-поисковая система
2	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
3	Беляева Л. Н. Лексикографический потенциал параллельного корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2004». СПб., 2004	текст, перевод, предложение	параллельный текст, машинный перевод, параллельный корпус
4	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
5	Гарабик Р., Захаров В. П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006	текст, выравнивание, предложение	словацкий национальный корпус, пользовательский интерфейс, морфологический разметка

Продолжение таблицы 2

№	Статья	Ключевые слова	Ключевые словосочетания
6	Зубов А. В. Корпус текстов белорусского языка // Труды международной конференции «Корпусная лингвистика – 2006». СПб., 2006	текст, корпус, кодирование	белорусский язык, корпус текстов, письменный текст
7	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
8	Копотев М. В. К построению частотной грамматики русского языка // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008	создание, корпус, Ханко	частотная грамматика, русский язык, частотные характеристики
9	Кустова Г. И. Электронный словарь степенной сочетаемости на базе Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008	словарь, значение, наречие	степенная сочетаемость, степенное слово, электронный словарь
10	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

4. 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; 12; 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 перефразировании, 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 перефразирования 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	Сергеева Е. М., Фивейская Е. А. Создание справочно-информационной базы данных для Лексического атласа русских народных говоров // Труды международной конференции «Корпусная лингвистика и лингвистические базы данных». СПб., 2002	Задача ЛАРНГ – показать в пространственной проекции (на карте) основные звенья словарного состава русских народных говоров, лексические и семантические различия в организации. Также в базе должна быть представлена следующая информация о картах, которые войдут в ЛАРНГ: данные об авторе-составителе; информация о материалах, полученных автором для составления карты; название карты (например, «ЛСЛ 123. Разработка такой базы данных входит в планы ИЛИ РАН на ближайшее время»	Создание справочно-информационной базы данных для Лексического атласа русских народных говоров (ЛАРНГ) является актуальной задачей русской диалектологии. Материалы будут иметь серьезное научное и культурно-историческое значение
2	Николаев И. С. Корпус текстов ижорских народных песен // Труды международной конференции «Корпусная лингвистика–2004». СПб., 2004	Работа с этим материалом показала, что для более полного сравнительно-исторического анализа ижорских диалектов необходимо изучить тексты, записанные у ижорцев ранее, а именно в конце XIX – начале XX вв. Несмотря на то, что в то время записей живой разговорной речи не велось, финскими учеными в Ингерманландии был собран обширный материал по устному народному творчеству, который	Корпус текстов ижорских народных песен в электронном виде был собран на кафедре математической лингвистики СПбГУ в рамках проекта «Полнотекстовая база данных по языкам и

Продолжение таблицы 3

№	Статья	Аннотация sumy	Аннотация rut5_base_sum_gazeta
2		был опубликован в многотомном собрании «Старые песни финского народа» [Suomen Kansan vanhat runot]. Решению всех этих и многих других проблем может способствовать создание корпуса текстов народных ижорских песен в электронном виде с соответствующей структурой и разметкой. Тем не менее, нам представляется, что, воспользовавшись опытом создания других корпусов текстов, можно решить перечисленные проблемы, а также те сложности, которые еще могут возникнуть при создании корпуса текстов ижорских народных песен	диалектам Северо-Запада России»
3	Котов А. А., Гопкало О. С. Русскоязычный эмоциональный корпус: коммуникативное взаимодействие в реальных эмоциональных ситуациях // Труды международной конференции «Корпусная лингвистика–2011». СПб., 2011	Эмоциональные корпуса важны для изучения общения с клиентами в состоянии стресса, для создания развлекательных технологий и для разработки эмоциональных компьютерных агентов: трехмерных компьютерных персонажей или роботов, способных взаимодействовать с человеком, распознавать его эмоции и правдоподобно имитировать собственные эмоции в процессе коммуникации. Для целей создания эмоциональных компьютерных агентов особый интерес представляют быстрые смены выражаемого эмоционального состояния и коммуникативных стратегий. Например, комбинация действий «поднимает брови» и «сжимает губы» (n = 47) может появляться в следующих контекстах: (a) Озадаченность: студент демонстрирует непонимание и озадаченность, смотрит в текст задания, поднимает брови, сжимает губы (20081219-zhum-a10, 20081227-fumo-a13), в некоторых случаях может при этом быстро моргать (20080717-c01, 20081227-fumo-a07)	В рамках проекта создания Русскоязычного эмоционального корпуса (REC) были фиксированы видеозаписи диалогов с клиентами в «службе одного окна» в одном из районов Москвы по вопросам оплаты коммунальных услуг
4	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 — рассматривается в качестве коллоката. Таким образом, при решении разных задач контекстной предсказуемости оказывается важным сопоставлять данные по сочетаниям как словоформ, так и лексем	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

5. 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; 20; 21; 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 4 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 4. 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	IntelliText, CAT&kittens, Revita, Текстометр, Visualizing Russian, 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	CWS (Chinese word segmentation), fastHan, LTP, PKUSeg, Ckiptagger
10	Корпусы	CORP	Русско-китайский параллельный корпус НКРЯ (ruzhcorp), НКРЯ, подкорпус RU-AC, CyberCAT, CAT, ruTenTen11
11	Тесты	TEST	Flesch Reading Ease, Flesch-Kincaid Grade

6. 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, предложенным челнами организационных комитетов конференций. Перед проведением экспертной разметки была осуществлена нормализация названий тем, результат представлен в таблице 5. Данная экспертная тематическая разметка будет использована для верификации результатов кластеризации текстов и тематических моделей, обученных на корпусе.

Таблица 5. 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

7. 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 RuTrmExtract и аннотации с помощью алгоритмов экстрактивной и абстрактивной суммаризации.

Планы дальнейшего развития проекта включают в себя проведение экспертной разметки ключевых выражений, уточнение формата генерируемых аннотаций, проведение процедур тематического моделирования и разработка поискового сервиса для работы с данными корпуса ТКиКЛ.

Литература

[1] Митрофанова О. А., Захаров В. П. 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: 09.02.2024).

[2] Виноградова Н. В., Митрофанова О. А. Формальная онтология как инструмент систематизации данных в русскоязычном корпусе текстов по корпусной лингвистике // Труды международной конференции «Корпусная лингвистика – 2008». СПб., 2008. С. 113–121.
URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (дата обращения: 09.02.2024).

[3] Виноградова Н. В., Митрофанова О. А., Паничева П. В. 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: 15.02.2024).

[4] Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб., 2020. 234 с.

[5] Тихонова Е. В., Косычева М. А. Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology. 2022. Вып. 3 (4). С. 7–15. URL: https://elibrary.ru/item.asp?id=49446588 (дата обращения: 12.03.2024).

[6] Kamshilova O., Beliaeva L., Geikhman L. 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). Saint Petersburg, Russia, November 27, 2019. CEUR Workshop Proceedings. 2020. Vol. 2552. P. 47–59. URL: https://elibrary.ru/item.asp?id=42584043 (дата обращения: 12.03.2024).

[7] Митрофанова О. А., Гаврилик Д. А. 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: 25.03.2024).

[8] Гусева Д. Д., Митрофанова О. А. 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. URL: https://human.spbstu.ru/userfiles/files/articles/2024/1/20-35.pdf (дата обращения: 26.03.2024).

[9] 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 / 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: 26.03.2024).

[10] Морозов Д. А. 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. С. 54–66. URL: https://elibrary.ru/lxeizh (дата обращения: 28.03.2024).

[11] Aries A., Zegour D., Walid H. 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: 29.03.2024).

[12] Nenkova A., McKeown K. Automatic summarization // Foundations and Trends in Information Retrieval. 2011. Vol. 5 (2–3). P. 103–233. URL: https://core.ac.uk/download/pdf/76383212.pdf (дата обращения: 30.03.2024).

[13] Allahyari M. 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. P. 397–405. URL: https://arxiv.org/abs/1707.02268 (дата обращения: 30.03.2024).

[14] Athugodage M., Mitrofanova O., Gudkov V. 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: 30.03.2024).
[15] Gudkov V., Mitrofanova O., Filippskikh E. 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: 30.03.2024).
[16] Pilault J. 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: 30.03.2024).
[17] Automatic text summarizer // PyPI. URL: https://pypi.org/project/sumy/ (дата обращения: 30.03.2024).
[18] RuT5SumGazeta // Hugging Face. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (дата обращения: 30.03.2024).
[19] Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. Recognizing Named Entities in Specific Domain // Lobachevskii Journal of Mathematics. 2020. Vol. 41 (8). P. 1591–1602. URL: https://link.springer.com/article/10.1134/S199508022008020X (дата обращения: 30.03.2024).
[20] Костюк Д. М., Широков Н. К. 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: 30.03.2024).
[21] Навроцкий А. А., Кривальцевич Е. В. 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: 30.04.2024).
[22] Yadav V., Bethard S. 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: 02.04.2024).
[23] Natasha // GitHub Repository. URL: https://github.com/natasha/natasha (дата обращения: 02.02.2024).
[24] Yargy // GitHub Repository. URL: https://github.com/natasha/yargy (дата обращения: 02.02.2024).
[25] Named Entity Recognition (NER) // DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (дата обращения: 02.02.2024).
[26] NEREL // GitHub Repository. URL: https://github.com/nerel-ds/NEREL (дата обращения: 02.02.2024).
[27] Stanford NER // David Batista. URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (дата обращения: 02.02.2024).

Text Corpus on Corpus Linguistics: Composition and Stages of Formation

O. A. Mitrofanova, M. A. Adamova, L. A. Bukreeva, A. K. Zernova,
A. A. Litvinova, V. S. Pavlikova, P. S. Sologub

Saint–Petersburg State University

The article is dedicated to the challenges of creating a corpus of articles on corpus 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 the text presentation format was carried out, and the structure of the articles was investigated. Experiments were conducted 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. An analysis was conducted on the distribution of conference reports into thematic blocks according to the expert annotation scheme.

Keywords: corpus linguistics, conference materials, annotation, keywords, summaries, thematic annotation, named entities

Reference for citation: Mitrofanova O. A., Adamova M. A., Bukreeva L. A., Zernova A. K., Litvinova A. A., Pavlikova V. S., Sologub P. S. Text Corpus on Corpus Linguistics: Composition and Stages of Formation // 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. 13–29. DOI: 10.17586/2541-9781-2024-8-13-29.

Reference

[1] Mitrofanova O. A., Zaharov V. P. Avtomatizirovannyj analiz terminologii v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Kompyuternaya lingvistika i intellektualnye tekhnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog 2009» (Bekasovo, 27–31 maya 2009 g.). Vyp. 8 (15). M.: RGGU, 2009. S. 321–328. URL: https://www.dialog-21.ru/digests/dialog2009/materials/pdf/49.pdf (access date: 09.02.2024). (In Russian)

[2] Vinogradova N. V., Mitrofanova O. A. Formalnaya ontologiya kak instrument sistematizacii dannyh v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Trudy mezhdunarodnoj konferencii «Korpusnaya lingvistika – 2008». SPb., 2008. S. 113–121. URL: https://project.phil.spbu.ru/corpora2011/Works2008/MitrofanovaVinogradova_113_121.pdf (access date: 09.02.2024). (In Russian)

[3] Vinogradova N. V., Mitrofanova O. A., Panicheva P. V. Avtomaticheskaya klassifikaciya terminov v russkoyazychnom korpuse tekstov po korpusnoj lingvistike // Trudy devyatoj Vserossijskoj nauchnoj konferencii «Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollekcii» (RCDL–2007). Pereslavl-Zalesskij, 2007. URL: http://rcdl.ru/doc/2007/paper_31_v1.pdf (access date: 15.02.2024). (In Russian)

[4] Zaharov V. P., Bogdanova S. YU. Korpusnaya lingvistika. SPb., 2020. 234 s. (In Russian)

[5] Tihonova E. V., Kosycheva M. A. Effektivnye klyuchevye slova: strategii formulirovaniya // Health, Food & Biotechnology. 2022. Vyp. 3 (4). S. 7–15. URL: https://elibrary.ru/item.asp?id=49446588 (access date: 12.03.2024). (In Russian)

[6] Kamshilova O., Beliaeva L., Geikhman L. 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). Saint Petersburg, Russia, November 27, 2019. CEUR Workshop Proceedings. 2020. Vol. 2552. P. 47–59. URL: https://elibrary.ru/item.asp?id=42584043 (access date: 12.03.2024).

[7] Mitrofanova O. A., Gavrilik D. A. Eksperimenty po avtomaticheskomu vydeleniyu klyuchevyh vyrazhenij v stilisticheski raznorodnyh korpusah russkoyazychnyh tekstov // Terra Linguistica.

2022. Vyp. 13 (4). S. 22–40. URL: https://elib.spbstu.ru/dl/2/j23-158.pdf/en/info (access date: 25.03.2024). (In Russian)

[8] Guseva D. D., Mitrofanova O. A. Klyuchevye vyrazheniya v russkoyazychnyh nauchno-populyarnyh tekstah: sravnenie vospriyatiya ustnoj i pismennoj rechi s rezultatami avtomaticheskogo analiza // Terra Linguistica. 2024 Vyp. 15 (1). S. 20–35. URL: https://human.spbstu.ru/userfiles/files/articles/2024/1/20-35.pdf (access date: 26.03.2024). (In Russian)

[9] 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 / 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 (access date: 26.03.2024).

[10] Morozov D. A. i dr. Generaciya klyuchevyh slov dlya annotacij russkoyazychnyh nauchnyh statej / Morozov D. A., Glazkova A. V., Tyutyulnikov M. A., Iomdin B. L. // Vestnik NGU. Seriya: Lingvistika i mezhkulturnaya kommunikaciya. 2023. № 1. S. 54–66. URL: https://elibrary.ru/lxeizh (access date: 28.03.2024). (In Russian)

[11] Aries A., Zegour D., Walid H. 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 (access date: 29.03.2024).

[12] Nenkova A., McKeown K. Automatic summarization // Foundations and Trends in Information Retrieval. 2011. Vol. 5 (2–3). P. 103–233. URL: https://core.ac.uk/download/pdf/76383212.pdf (access date: 30.03.2024).

[13] Allahyari M. 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. P. 397–405. URL: https://arxiv.org/abs/1707.02268 (access date: 30.03.2024).

[14] Athugodage M., Mitrofanova O., Gudkov V. 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/ (access date: 30.03.2024).

[15] Gudkov V., Mitrofanova O., Filippskikh E. 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/ (access date: 30.03.2024).

[16] Pilault J. 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/ (access date: 30.03.2024).

[17] Automatic text summarizer // PyPI. URL: https://pypi.org/project/sumy/ (access date: 30.03.2024).

[18] RuT5SumGazeta // Hugging Face. URL: https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta (access date: 30.03.2024).

[19] Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. 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 (access date: 30.03.2024).

[20] Kostyuk D. M., Shirokov N. K. Metody identifikacii imenovannyh sushchnostej v zadachah obrabotki potoka nauchnyh novostej // Menedzhment vuzovskih bibliotek. Minsk, 2021. S. 50–54. URL: https://elibrary.ru/item.asp?id=49171334 (access date: 30.03.2024). (In Russian)

[21] Navrockij A. A., Krival'cevich E. V. Sravnitel'nyj analiz sistem izvlecheniya imenovannyh sushchnostej iz nestrukturirovannyh publicisticheskih tekstov // BIG DATA and Advanced Analytics = BIG DATA i analiz vysokogo urovnya. Minsk, 2020. S. 12–18. URL: https://elibrary.ru/item.asp?id=43934323 (access date: 30.04.2024). (In Russian)
[22] Yadav V., Bethard S. 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 (access date: 02.04.2024).
[23] Natasha // GitHub Repository. URL: https://github.com/natasha/natasha (access date: 02.02.2024).
[24] Yargy // GitHub Repository. URL: https://github.com/natasha/yargy (access date: 02.02.2024).
[25] Named Entity Recognition (NER) // DeepPavlov. URL: https://docs.deeppavlov.ai/en/master/features/models/NER.html (access date: 02.02.2024).
[26] NEREL // GitHub Repository. URL: https://github.com/nerel-ds/NEREL (access date: 02.02.2024).
[27] Stanford NER // David Batista. URL: https://www.davidsbatista.net/blog/2018/01/23/StanfordNER/ (access date: 02.02.2024).