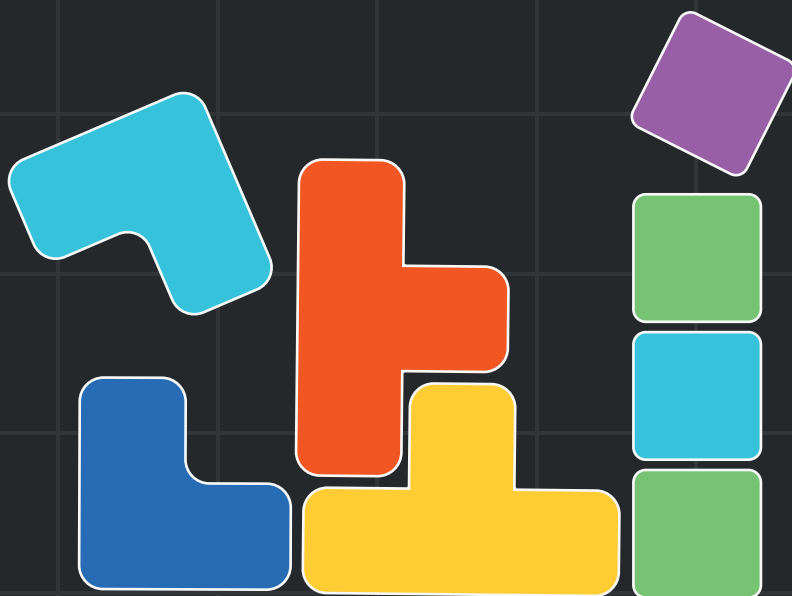


IDO / 2025
CONF

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА (NLP) ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА

Сборник материалов международной
научно-практической конференции
«Обработка естественного языка (NLP)
для лингвистики, IT и бизнеса»



28 февраля — 1 марта 2025
Томск

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА

**Сборник материалов
Международной научно-практической конференции
(28 февраля – 1 марта 2025 г.)**

Томск
Издательство Томского государственного университета
2025

УДК 004.8
ББК 32.813
О13

Редакционная коллегия:

З.И. Резанова (отв. ред.); А.В. Бочаров; К.П. Крапивина

О13 Обработка естественного языка для лингвистики, IT, бизнеса :
сборник материалов Международного научно-практического
семинара (30 февраля – 1 марта 2025 г.) / отв. ред.
З.И. Резанова. – Томск : Издательство Томского
государственного университета, 2025. – 58 с.

ISBN 978-5-907890-87-9

В сборнике представлены результаты научных исследований и прикладных проектов, реализуемых с использованием методов и технологий автоматической обработки естественного языка (NLP). Основные направления научных и прикладных разработок: создание и использование корпусов текстов, баз данных, чат-ботов; автоматизированный анализ текстов с применением когнитивных технологий; использование больших языковых моделей для генерации и обработки текстов.

Для исследователей и практиков, заинтересованных в разработке и применении технологий NLP и AI для решения широкого круга социально значимых проблем.

УДК 004.8
ББК 32.813

Рецензенты:

А.В. Колмогорова, доктор филологических наук, профессор;
К.С. Шиляев, кандидат филологических наук, доцент

ISBN 978-5-907890-87-9 © Томский государственный университет, 2025

Предисловие

В сборнике представлены тезисы участников Международного научно-практического семинара «Обработка естественного языка (NLP) для лингвистики, IT, бизнеса», организованного Томским государственным университетом феврале-марте 2025 г.

Семинар первоначально задумывался как организационно-научная площадка для представления и обсуждения научных исследований и прикладных проектов магистрантов, обучающихся на сетевой магистерской программе «Обработка естественного языка для лингвистики и IT» в Томском государственном университете. Миссию программы ее создатели видят в том, чтобы готовить специалистов в сфере NLP, способных включаться в высокотехнологичные быстро развивающиеся проекты индустриальных и исследовательских команд. Базовая идея, на которой строится программа, – интеграция науки и бизнеса в преподавании академических дисциплин и при реализации выпускных квалификационных проектов магистрантов. Мы высоко ценим участие в реализации программы ведущих специалистов предприятий IT-индустрии АйПиЭсТи (<https://ips-tomsk.ru/>), Эксмо (<https://eksmo.ru/>), Палекс (<https://palexgroup.com/>), Образовариум (<https://obr.nd.ru/>), Skillfactory (<https://skillfactory.ru/>). Проекты, выполненные на базе, по заказу и при участии специалистов данных компаний, были успешно представлены на конференции.

Однако уже на этапе планирования фокус и амбиции организаторов существенно изменились и мероприятие было проведено в формате международного научно-практического семинара с привлечением ведущих специалистов академического сообщества наравне с представителями IT-индустрии, активно использующими и разрабатывающими идеи и технологии NLP.

На пленарных докладах конференции были представлены обобщающие концепции использования NLP в науке и технологиях и конкретные исследовательские разработки. Так, в докладах А.Г. Крайнова, директора по развитию технологий искусственного интеллекта Яндекс (<https://yandex.ru/company/>) и Р. Дас (Rajib Das), Sourav Mukherjee, Avijit Pyne, руководителей компании Bayshore Intelligence Solutions (<https://bayshoreintel.com/>), лидера на IT-рынке Индии, были представлены две интерпретации масштабного включения NLP-технологий в создание продуктов современных высокотехнологических рынков. В докладе В. Васильева, руководителя проектов единых рекомендаций AI VK (<https://team.vk.company/>), наряду с аспектом заказа к NLP со стороны бизнеса обсуждался вопрос о применении методов NLP при решении исследовательских и социальных задач. В.А. Малых, руководитель NLP исследований МТС ИИ (<https://mts.ai/>) представил исследовательскую разработку «Итеративное самообучение для генерации кода с помощью усиленного повторного ранжирования (Iterative Self-Training for Code Generation via Reinforced Re-Ranking). О. Кольцова, д.ф.н., директор Лаборатории социальной и когнитивной информатики НИУ ВШЭ в Санкт-Петербурге (<https://scila.hse.ru/>) охарактеризовала перспективы использования больших языковых моделей для лингвистики и общества. Коллеги из Школы лингвистики НИ ВШЭ (<https://ling.hse.ru/>) проф. Е.В. Рахилина, доц. Д.А. Рыжова, доц. А.С. Выренкова представили компьютерные проекты Школы лингвистики Высшей школы экономики. Мы благодарны пленарным докладчикам за участие в качестве модераторов и экспертов на секционных заседаниях

В разделах сборника представлены в соответствии с тематическим делением секций тезисы докладов участников секционных заседаний конференции.

В секции «Корпуса и базы данных для исследовательских задач» обсуждались общие концепции и технологии создания корпусов текстов, конкретные теоретические и технологические решения при моделировании корпусов, а также варианты использования корпусных данных в теоретических исследованиях и прикладных проектах.

Материалы участников секции «Интеллектуальный анализ текстов для социально-значимых задач» представляют конкретные решения задач, находящихся в двух проблемных областях: методы определения сложности текстов и аспекты автоматического выделения аспектов семантики, сентимента текстов разных дискурсов и жанров.

Участники секции «Языковые модели в генерации и обработке текстов» в материалах сборника представляют направления использования технологий искусственного интеллекта к решению широкого круга задач применительно к разным языкам (русский, английский, тибетский, арабский), в пределах последних – к разным дискурсам (новостному, художественному, научному).

Авторы докладов секции «Промпт-инжиниринг, разработка чат-ботов и прочие прикладные NLP-задачи» представляют прикладные проекты создания чат-ботов, используемых в образовательных и просветительских проектах, а также обсуждают другие аспекты использования технологий искусственного интеллекта в разных отраслях и способы его оптимизации.

Организационный комитет конференции первого Международного научно-практического семинара «Обработка естественного языка (NLP) для лингвистики, IT, бизнеса» выражает благодарность всем участникам конференции за создание атмосферы научного поиска и плодотворных дискуссий на семинаре 2025 г. и приглашают к дальнейшему сотрудничеству исследователей и практиков, активно работающих в сфере NLP.

З.И. Резанова, проф., д-р филол. наук

КОРПУСА И БАЗЫ ДАННЫХ ДЛЯ ИССЛЕДОВАТЕЛЬСКИХ ЗАДАЧ

РАЗРАБОТКА СЛОВАРЯ-ТЕЗАУРУСА МИСТИКО-АСКЕТИЧЕСКОЙ ТРАДИЦИИ ИСИХАЗМА

Э.А. Биденко

Томский государственный университет, Томск, Россия

В рамках стратегической исследовательской задачи по созданию универсальной модели аргументированных рассуждений для различных предметных областей решается подзадача моделирования базы знаний. В качестве предметной области выбрана мистико-аскетическая традиция исихазма, что обусловлено богатым наследием святоотеческой письменности, значительным вкладом научной школы «Синергийной антропологии» и уникальностью исихастского материала, характеризующегося закрытостью дискурса, маргинальностью традиции и внутренним мистическим опытом.

Для структурной метаразметки словаря-тезауруса использован «Аналитический словарь исихазма», представленный профессором Хоружим С.С. в работе «К феноменологии аскезы» (1998). Метаразметка включает четыре уровня абстракции: группы (5), категории (24), субкатегории и энергии.

В ходе исследования реализован интерфейс приложения СУБД на архитектуре PostgreSQL с графопостроителем GraVis. Проведена адаптация словарных источников к реляционной архитектуре БД, включая Большой толковый словарь русского языка С.А. Кузнецова, Толковый словарь В.И. Даля и Словарь-

тезаурус эмотивной лексики Л.Г. Бабенко. Разработан начальный проект реляционной БД словаря-тезауруса исихазма. Подготовлен корпус текстов объемом 381 текст (в среднем 115 предложений) из источника «Добротолюбие» в 5 томах. Создан аннотированный корпус из 2400 предложений для обучения классификаторов по 5 группам и 24 категориям исихазма. Обучены мультиклассификаторы для групп situation, motivation, praxis, feoria, eschatology и для соответствующих категорий с достигнутой точностью от 0.65 до 0.99.

Перспективы дальнейших исследований включают разработку БД дескрипторной метафорической модели исихастского дискурса и методов автоматического извлечения тезаурусных отношений из текста. Практическое внедрение результатов предусматривает использование словаря-тезауруса как справочного источника знаний для модели аргументированных рассуждений, позволяющего реализовать управляемую генерацию аргументативных высказываний. Также планируется применение разработки для решения образовательных задач в области нравственного и сравнительного богословия, сектологии и миссионерской деятельности, а также автоматизации компаративного подхода в исследовании духовных практик.

СОСТАВЛЕНИЕ КОРПУСА РИТОРИЧЕСКИХ СТРУКТУР АНГЛОЯЗЫЧНОЙ И РУССКОЯЗЫЧНОЙ ПЕСЕННОЙ ПОЭЗИИ И ЕГО ВОЗМОЖНЫЕ ПРИМЕНЕНИЯ

А.В. Зоркальцев

Новосибирский государственный университет, Новосибирск, Россия

Теория риторических структур (ТРС) – способ репрезентации семантики и прагматики текста, позволяющий представить текст

в виде дерева смысловых отношений между описываемыми в тексте ситуациями. Изначально эту теорию планировалось применять в обработке естественных языков. Преимущество ТРС в данной области – это осмысленность: речь идёт не о поверхностных статистических данных о встречаемости слов, а о математических объектах, отражающих смысловую организацию текста. Однако подготовка достаточно большого корпуса риторических структур – очень затратный процесс. Поэтому данную теорию стали использовать в анализе дискурса, когнитивных, стилеметрических и иных лингвистических исследованиях.

Целью нашего исследования было проверить применимость ТРС к песенной поэзии как к наиболее доступной форме литературы. Одной из главных задач было составление корпуса. Требовалось выдвинуть критерии к отбору текстов (художественная ценность в отрыве от музыки; техническая отделка и др.). Сначала корпус включал тексты русских поэтов-песенников, представителей разных направлений: Рождественского, Городницкого, Шевчука (группа «ДДТ») и Мозжухина (группа «Дайте Танк (!)»). Работая с этим материалом, мы также ввели сложностные и эстетические метрики, рассчитываемые по математическим характеристикам дерева риторических структур (количество листьев, высота, частотный состав отношений и др.) и определяющие, насколько текст является сложным или эстетичным. Метрики были проверены с помощью анкетирования. Некоторые из них оказались достаточно эффективными.

Затем корпус был дополнен англоязычными текстами Дилана (американский английский), Маккартни (британский английский), Кейва (австралийский английский) и Коэна (канадский английский). Сейчас он включает 80 текстов. Для разметки использовался набор из 34 риторических отношений. Всего корпус включает 2372 идентифицированных отношения (что приблизительно равно количеству пропозиций в корпусе). Исследование

имеет следующие перспективы: коррекция некоторых полученных значений метрик, зависящих от вероятностей появления риторических отношений; постановка стилеметрического эксперимента (например, такого: определить авторство текста, используя данные корпуса); использование корпуса для задач машинного обучения.

КОРПУСНЫЕ ТЕХНОЛОГИИ В ИЗУЧЕНИИ ТВОРЧЕСТВА Ф.М. ДОСТОЕВСКОГО

Е.М. Понкратова

Томский политехнический университет, Томск, Россия

Творчество Фёдора Михайловича Достоевского остаётся одной из важнейших составляющих мировой культуры, продолжая оказывать глубокое воздействие на литературу, философию и общественные науки. Его произведения отличаются сложностью языка, многогранностью образов и идей, что делает их идеальным объектом для применения новейших научных подходов.

В последние десятилетия развитие компьютерных технологий открыло перед учёными новые горизонты для изучения языка и стиля писателей, предоставив инструменты для объективного и глубокого анализа текстов.

Корпусная лингвистика представляет собой одно из наиболее перспективных направлений в этой области. Её методы позволяют проводить количественный и качественный анализ больших массивов текстов, выявляя скрытые закономерности. Применение корпусных технологий даёт возможность исследовать творчество Достоевского с новой стороны. Среди методов корпусной лингвистики, применяемых в исследованиях творческого наследия Достоевского выделяются следующие:

Частотный анализ ключевых слов и тем. Исследователи используют корпусные методы для определения наиболее часто употребляемых слов и фраз в произведениях Достоевского, что помогает выявить основные темы и идеи его творчества. Создан «Статистический словарь языка Достоевского».

Анализ стиля и языка. Корпусные методы применяются для изучения стиля и языковых особенностей произведений Достоевского, включая выбор слов, синтаксические конструкции и другие языковые аспекты. Институтом русского языка им. В. В. Винаградова создан значительный труд «Корпусная модель идиостиля Достоевского», который показывает существенные особенности идиостиля писателя. Кроме того специалистами Института системного программирования РАН была создана интернет-платформа для работы с частностью предъявлений определённых слов и выражений, поиска идиом в различных корпусах текстов;

Создание конкордансов. Исследователи творчества писателя из Петрозаводского государственного университета создали конкордансы художественных произведений Достоевского – списки всех случаев употребления определённых слов или фраз в авторской орфографии.

Контекстуальный и тематический анализ, метод системного статистического анализа всего массива текстов писателя (с 1838 по 1881 г.) позволили исследователям Томского государственного университета реконструировать геополитическую карту Ф.М. Достоевского. В результате проведённого исследования было выявлено 55 геополитических субъектов, определено общее количество упоминаний: 8278, уточнён список художественных и публицистических произведений, в которых писатель обращается к геополитической проблематике. Итогом проделанной работы стало создание многоуровневой интерактивной карты, созданной с использованием методов корпусной лингвистики при анализе материалов творчества Достоевского.

СОЗДАНИЕ КОРПУСА DOSPENA ДЛЯ РАСПОЗНАВАНИЯ HATE SPEECH НА МАТЕРИАЛЕ ЧАТА ВИДЕОИГРЫ DOTA 2

Л.А. Сушков¹, Е.П. Трунцева¹

¹ Томский государственный университет, Томск, Россия

Задача нашей работы состоит в автоматическом распознавании Hate Speech в чате видеоигр. Ввиду нехватки языкового материала по теме, мы решили составить свой корпус для решения данной задачи.

На первом этапе методом парсинга мы собрали сообщения из открытого источника OpenDota и отобрали только русскоязычные сообщения. На втором этапе мы составили инструкцию для разметки сообщений по трем аспектам – степень языка вражды «Hate Speech» (тэги 0–3), наличие обценной лексики «Profanity» (тэги 0–1), связь с прошлыми сообщениями «Relates to previous» (тэги 0–1). На третьем этапе тремя аннотаторами было размечено 35 526 сообщений, после чего была посчитана согласованность разметки между ними.

Используя коэффициент Каппа Флейсса для оценки согласованности мы получили следующие результаты по каждому из аспектов: Hate Speech – 70,3%, Profanity – 90,7%, Relates to previous – 11,4%. Затем, мы объединили разметку трех аннотаторов по каждому из аспектов и получили корпус, состоящий из трех данных – текст сообщений, тэг Hate Speech, тэг Profanity. Получившийся корпус имеет несбалансированные классы, где один аспект имеет явное количественное преимущество над другими (Hate Speech, тэг 0–29547 сообщений; Profanity, тэг 0–28374 сообщений).

В дальнейшем полученный корпус планируется использовать для создания моделей автоматической классификации по аспектам Hate Speech и Profanity. Также мы планируем сделать данный корпус публичным для других исследователей по данной тематике.

СОЗДАНИЕ КОРПУСА ЗВУЧАЩЕЙ РЕЧИ ДЛЯ ЛИНГВИСТИЧЕСКИХ И МЕЖДИСЦИПЛИНАРНЫХ ИССЛЕДОВАНИЙ

У.Э. Чекмез

Иркутский государственный университет, Иркутск, Россия

На кафедре русского языка и общего языкознания ИГУ реализуется проект «Устный корпус как инструмент лингвистических и междисциплинарных исследований». Главная цель проекта – создание общедоступного корпуса звучащей речи на базе материалов, собранных в рамках предыдущих социолингвистических и диалектологических проектов кафедры («Устная история Иркутска», «Русские говоры Прибайкалья в синхронии и диахронии»).

Работа с материалами корпуса проходит в несколько этапов:

– на первом этапе с помощью программы ELAN фиксируются речевые и неречевые элементы вокального канала, происходит деление потока звучащей речи на элементарные дискурсивные единицы и их временная привязка, данные выгружаются в формате TextGrid;

– на втором этапе с помощью программы Praat осуществляется ручная разметка просодии, после чего файлы конвертируются в формат JSON для упрощения работы с языками программирования, используемыми при создании сайта;

– на третьем этапе осуществляется морфологический анализ вербальных единиц с помощью библиотеки Stanza, его ручная корректировка и запись этих данных в созданный на предыдущем этапе файл JSON;

– на четвертом этапе предполагается ручная разметка тематики нескольких текстов экспертами, полученный набор данных выступит материалом для обучения модели RoBERTa и дальнейшей автоматической разметки тематики текстов.

Результатом работы станет общедоступный корпус звучащей речи, состоящий из двух подкорпусов («Устная история Иркутска» и «Русские говоры Прибайкалья»). В корпусе предполагается возможность организации лексико-грамматического поиска и поиска по просодии, обеспечивающего возможность лингвистических исследований устной речи, и тематического поиска, который может представлять интерес для специалистов других областей гуманитарного знания. Созданием сайта и разработкой системы поиска занимаются программисты ИМИТ ИГУ, на сегодняшний день сайт существует в альфа-версии и доступен только для участников проекта.

АНАЛИТИКА УЧЕБНЫХ ТЕКСТОВ ДЛЯ ИЗУЧЕНИЯ РУССКОГО ЯЗЫКА КАК ИНОСТРАННОГО: ФУНКЦИОНАЛ ПЛАТФОРМЫ RULINGVA*

М.И. Андреева¹, М.И. Солнышкина¹, А.В. Данилов¹

¹ Казанский (Приволжский) федеральный университет, Казань, Россия

Профилирование и классификация текстов, контрастирование их лингвистических параметров с целью выявления референсных диапазонов для последующего структурирования больших

* Исследование выполнено за счет гранта Российского научного фонда № 24-78-10129, <http://rscf.ru/project/24-78-10129/>.

языковых данных рассматривается в настоящее время как весьма актуальная задача в текстовой аналитике в частности и прикладной лингвистике в целом. Выполнение такого рода задач возможно только при помощи мощных текстовых анализаторов и профайлеров, широко используемых редакторами, разработчиками тестовых и учебных материалов, исследователями-лингвистами.

Представленное исследование осуществлено с целью выявления относительной сложности двух текстов уровня А2, используемых в практике преподавания русского языка как иностранного (далее РКИ), и выполнено с использованием текстового анализатора RuLingva (rulingva.kpfu.ru/lexdiv). RuLingva рассчитывает метрики 120 параметров текстов на русском языке, включая уровень сложности (А1–С2) каждой лексемы и может выступать в качестве инструмента, способного прогнозировать трудность понимания текста для респондентов разных категорий. Материалом исследования выступили адаптированные для изучения русского языка как иностранного (далее РКИ) тексты «Злой мальчик» и «Месть» (по А.П. Чехову), объем которых составляет 341 и 576 словоформ соответственно. Тексты маркированы А2_1 и А2_2. В качестве источника использован Учебный корпус русского языка (Свидетельство о регистрации базы данных RU 2020622254).

Отдельную проблему в преподавании и тестировании РКИ представляет ранжирование текстов по уровням. В изучаемых текстах А2 доля лексики уровня А2 составляет только 3,23% (А2_1) и 5,38% (А2_2), а лексика уровня В1 – 8,50% и 11,98%. Около 6% лексики в каждом из текстов относятся к уровню В2. Параметрический анализ текстов также выявил 20 (7,33%) и 34 (5,90%) лексемы, отсутствующие в лексических минимумах РКИ. Большую часть данных слов составляют именованные сущности, а именно, имена, отчества и фамилии персонажей. Например, *Анна, Семёновна, Иванович, Лапкин, Саввич, Турманов, Дегтярёв и др.*

RuLingva представляет пользователям возможность осуществлять синонимические замены в тексте, если текст содержит лексические единицы более высоких уровней. Например, слово *отличный*, очевидно, не знакомое студентам уровня A2, предлагается заменить на *хороший, идеальный, прекрасный*, присутствующие в списках A2. Приведем примеры синонимов, использованных при адаптации оригинального текста «Месь» (Чехов А.П.) для уровня A2 РКИ. Определение уровня сложности лексических единиц автоматизировано на платформе RuLingva и указывается в скобках рядом с каждым словом. Пример, «Пожалуй (C2), пожалуй (C2) ...» (Оригинал) – «Конечно (A1), конечно (A1)» (A2_РКИ); «густой (B2) сочный (B2) бас (B2)» (Оригинал) – «мужской (A1) голос (B1)» (A2_РКИ).

Сложность лексики текстов

Текст	Текст РКИ до RuLingva	Текст РКИ после замены синонимов на RuLingva
A2_1	Иван (A1) Иванович (UNL) Лапкин (UNL), <u>приятный</u> (B1) молодой (A1) человек (A1), и (A1) Анна (A1) Семёновна (UNL), молодая (A1) девушка (A1), пришли (A1) на (A1) берег (A2) реки (A1), <u>спустились</u> (B2) к (A2) воде (A1) и (A1) сели (A2) на (A1) скамейку (B2)	Иван (A1) Иванович (UNL) Лапкин (UNL), <u>хороший</u> (A1) молодой (A1) человек (A1), и (A1) Анна (A1) Семёновна (UNL), молодая (A1) девушка (A1), пришли (A1) на (A1) берег (A2) реки (A1), <u>подошли</u> (B1) к (A2) воде (A1) и (A1) сели (A2) на (A1) скамейку (B2)
A2_2	Но (A1) ему (A1) было (A1) очень (A1) <u>неприятно</u> (B2) и (A1) <u>обидно</u> (B2), что (A1) его (A1) <u>приятель</u> (B2) называет (A1) его (A1) старым (A1) толстым (B1) <u>дураком</u> (C1)	Но (A1) ему (A1) было (A1) очень (A1) <u>плохо</u> (A1) и (A1) <u>грустно</u> (A2), что (A1) его (A1) <u>друг</u> (A1) называет (A1) его (A1) старым (A1) толстым (B1) <u>глупым</u> (A2) человеком (A1)

В качестве дополнительных функций RuLingva предлагает разбивку на предложения, их нумерацию, визуализацию различных метрик, частотность отдельных слов и др.

ОТ ПОЛЕВЫХ ДАННЫХ К АВТОМАТИЧЕСКОМУ АНАЛИЗУ: РАЗРАБОТКА МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ НАУКАНСКОГО ЯЗЫКА*

**Е.М. Будянская¹, А.О. Бузанов^{1,2},
Д.О. Жорник¹, А.А. Пихтин^{1,2}**

¹ Институт языкознания РАН, Москва, Россия

² Высшая школа экономики, Москва, Россия

В сфере компьютерной обработки естественного языка принято разделять языки на много- и малоресурсные. Многоресурсные языки, такие как английский, испанский или русский, характеризуются наличием обширных корпусов текстов, размеченных датасетов и развитых инструментов обработки. К малоресурсным относятся языки с ограниченным количеством цифровых ресурсов – такие как албанский, грузинский, исландский или хинди. Особую категорию составляют экстремально малоресурсные языки – преимущественно языки без государственного статуса, для которых электронные ресурсы практически отсутствуют. В эту категорию попадает большинство языков России.

В конце XX в. индустрия естественной обработки языка совершила прорыв, перейдя от правилых методов к статистиче-

* Работа ведется в рамках проекта по документации науканского языка, поддержанного грантом РНФ № 23-18-00204 «Динамика языковой ситуации в сообществах Крайнего Севера: диахроническая документация науканского языка» (рук. В.А. Плунгян).

ским. Однако для подавляющего большинства языков мира применение статистических методов остается проблематичным из-за недостатка данных. В таких случаях по-прежнему необходимо использовать либо правилые модели для автоматической разметки, либо прибегать к трудоёмкой ручной разметке данных.

Наше исследование фокусируется на разработке морфологического анализатора для науканского языка (< юпикские < эскимосские < эскимосско-алеутские). Создание правилового морфологического анализатора для полисинтетического языка с высокой морфологической сложностью требует значительных усилий по формализации грамматических правил. Однако такой подход имеет существенное преимущество: полное описание морфологии позволяет автоматически (с определёнными ограничениями) выделять корни слов.

Учитывая, что лексика представляет собой значительно более обширный и, что важно, открытый класс по сравнению с грамматикой, мы предлагаем подход к формированию словаря по принципу “всё кроме грамматики”. Этот метод позволяет эффективно использовать морфологический анализатор для автоматического пополнения словаря.

В нашем докладе представлено детальное сравнение двух подходов: полноценного морфологического анализатора, использующего словарь, и морфологического гессера, работающего без информации о корнях науканского языка. Результаты нашего исследования демонстрируют, что при работе с полевыми данными использование морфологического гессера может быть более эффективным, чем применение полного анализатора, несмотря на известные ограничения. Это особенно актуально на начальных этапах документации языка, когда словарь еще находится в процессе формирования.

СОВРЕМЕННЫЕ МЕТОДЫ ИСПОЛЬЗОВАНИЯ РАЗМЕТКИ ИМЕНОВАННЫХ СУЩНОСТЕЙ НА ПРИМЕРЕ РАЗМЕТКИ КОРПУСА ДЕТСКОЙ ЛИТЕРАТУРЫ ИЗ КАТАЛОГА БИБЛИОТЕКИ МАРТЬЯНОВА

М.С. Брок

Высшая школа экономики, Санкт-Петербург, Россия

Работа посвящена современным методам разметки именованных сущностей (NER) с акцентом на детскую литературу из каталога библиотеки Мартьянова. Идентификация и классификация именованных сущностей являются важными задачами в области компьютерной лингвистики и обработки естественного языка, применяемыми в таких сферах, как машинный перевод и анализ текстов. В работе исследуется специфика применения современных методов NER к детской литературе.

Для анализа были рассмотрены приключенческие романы из каталога детской литературы. В ходе исследования был выявлен эффективный подход к разметке с использованием библиотеки *Natasha*, которая показала высокую точность в идентификации именованных сущностей. Результаты работы могут служить основой для создания специализированных предобученных моделей, способных учитывать языковые особенности детской литературы и улучшить результаты автоматической обработки текстов в данной области. В дальнейшем планируется исследовать возможность адаптации существующих моделей для более точной обработки таких текстов.

Актуальность исследования. Современные задачи обработки естественного языка (NLP) требуют высокоточной иденти-

фикации и классификации именованных сущностей (NER). Детская литература обладает специфическими языковыми особенностями, такими как упрощенная грамматика, образные имена, наличие вымышленных персонажей, что усложняет применение традиционных методов NER. Недостаточная представленность детской литературы в существующих корпусах затрудняет разработку методов их автоматической обработки в данной категории. Однако использование специализированных подходов к разметке именованных сущностей в детской литературе может значительно улучшить качество обработки текстов для образовательных и библиотечных систем.

Область применения. Разметка именованных сущностей в детской литературе востребована в таких сферах, как машинный перевод, автоматическая каталогизация и поиск книг в библиотеках, в лингвистических исследованиях и анализе текстов. Результаты работы призваны способствовать развитию специализированных предобученных моделей для обработки детских текстов, что повысит точность работы систем обработки естественного языка в вышеназванных областях.

Объект исследования. Объектом исследования был выбран корпус приключенческих романов из каталога детской книги библиотеки Мартянова в городе Минусинске. Особое внимание было уделено изучению языкового стиля и структуры именованных сущностей в детской литературе, что позволило выявить особенности обработки таких текстов в рамках NER.

Ход исследования и результаты. В рамках исследования проведен обзор некоторых современных методов NER (Spacy, NLTK Natasha), а также разметка текстов детской литературы с применением библиотеки Natasha. Проведена оценка качества разметки и анализ эффективности инструментов NER в данной предметной области, которые позволили выявить наиболее эффективный подход для работы с детскими текстами.

Анализ показал, что Natasha наиболее эффективно справляется с задачей разметки именованных сущностей в детской литературе, обеспечивая высокую точность идентификации именованных сущностей. Были выявлены особенности именованных сущностей в детских текстах, среди которых преобладание антропонимов, а также наличие вымышленных топонимов и организаций. Обоснована необходимость создания специализированных моделей NER, адаптированных к языковым особенностям детских текстов

Перспективы. В дальнейшем планируется разработка и дообучение специализированных моделей NER на корпусах детской литературы. Ожидается, что применение результатов исследования позволит улучшить автоматический анализ и обработку текстов в образовательных и библиотечных системах, а также повысить качество работы с детской литературой в системах машинного перевода и интеллектуального поиска.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ ДЛЯ СОЦИАЛЬНО ЗНАЧИМЫХ ЗАДАЧ

СЛОЖНОСТЬ ТЕКСТА КАК МЕЖДИСЦИПЛИНАРНАЯ КАТЕГОРИЯ*

Е.В. Мартынова¹, Р.В. Куприянов¹, М.И. Солнышкина¹

¹ Казанский (Приволжский) федеральный университет, Казань, Россия

Представители различных научных школ, включая российскую, американскую, немецкую, румынскую и проч., трактуют сложность текста как конструкт, включающий лингвистическую и когнитивную составляющие. Парадигма исследований данной проблематики располагает несколькими терминами, обозначающими объект(ы) исследования: «сложность текста», «трудность текста», «понятность текста», «доступность текста», «читабельность», «читаемость», «удобочитаемость».

Лингвистическая сложность текста как мера сложностей составляющих текст единиц рассчитывается как аддитивная категория на основе метрик предикторов сложности. Предикторы сложности соотносятся с языковыми уровнями: морфологическим, лексическим, синтаксическим и дискурсивным. Для неалфавитных языков дополнительно к указанным оцениваются, в

* Исследование выполнено при поддержке гранта Российского научного фонда, проект № 24-28-01355 «Жанрово-дискурсивные характеристики текста как функция лексического диапазона».

частности, разнообразие иероглифов, структурная сложность иероглифов, частотность иероглифов, психолингвистическая информация о восприятии иероглифов. Предикторы каждого уровня сложности валидируются экспериментально на различных категориях респондентов.

Когнитивная сложность текста как совокупность квантов информации, содержащихся в тексте, традиционно рассчитывается при помощи пропозиционального анализа и трактуется как объективная категория. Информативность текста может варьироваться, т.е. быть индивидуальной, для отдельных читателей и зависима от их подготовленности и знаний.

Междисциплинарный подход позволяет интегрировать исследования данной проблематики в отдельную область – дискурсивную комплексологию, аккумулировавшую алгоритмы оценки трудности текста на основе анализа лингвокогнитивных способностей читателей, а сложности текста – при помощи экспертной оценки и автоматических профайлеров.

При оценке сложности текста каждая из указанных составляющих оценивается по двум основным шкалам, ранжирующим тексты, предназначенные для носителей языка, по годам обучения, а для изучающих язык как иностранный – по уровням владения, сформированным Общеввропейской шкалой (CEFR).

Алгоритм автоматической оценки лингвистической сложности текста или профилирование включает следующие этапы: (1) формирование коллекции языкового материала; (2) метаописание документов коллекции; (3) предобработка, т.е. удаление фрагментов, контаминирующих однородность типа/жанра/регистра/функционального стиля; (4) расчет значений/метрик параметров-предикторов сложности; (5) установление степени сложности на основе карт референсных диапазонов значений параметров.

В качестве основополагающих положений дискурсивной комплексологии приняты следующие: (1) предикторы сложности

текста жанрозависимы; (2) формулы читабельности предлагают весьма общую, но допустимую в ряде практических ситуаций, оценку трудности восприятия текста читателем; (3) достоверность сравнительной оценки сложности текстов одного или разных жанров достигается при условии (а) единой для каждого из них читабельности и (б) последовательной оценки метрик их морфологических, лексических, синтаксических и дискурсивных параметров.

Исследовательские ниши современной дискурсивной комплексологии включают следующие: (1) списки параметров-предикторов сложности всех уровней сложности для текстов разных жанров; (2) списки параметров-дискриминантов текстов различных типов и жанров; (3) референсные диапазоны значений лингвистических параметров-предикторов каждого уровня сложности (2-11 классы, A1-C2, CEFR).

Результаты исследования дискурсивной комплексологии используются в различных прагмалингвистических ситуациях для прогнозирования успешности понимания текста читателями определенного возраста, года обучения, профессиональной специализации.

ПОДХОДЫ К ОЦЕНКЕ СЛОЖНОСТИ ТЕКСТОВ, ПРЕДНАЗНАЧЕННЫХ ДЛЯ ТЕСТИРОВАНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ

А.С. Засыпкин¹, Е.П. Бручес¹

¹ Новосибирский государственный университет, Новосибирск, Россия

Языковые модели стали частью повседневности, и тщательное их тестирование является гарантией их качества и надёжности.

Однако тексты, используемые для этого, в свою очередь также нуждаются в оценке. Один из аспектов, который важно учитывать – это сложность текста и её потенциальное влияние на качество работы моделей (а также соответствие сложности входного и выходного текста). В нашем исследовании мы проводим комплексный анализ факторов, влияющих на сложность текста, на материале популярных бенчмарков – наборов данных для тестирования языковых моделей.

Исследование включает в себя несколько ключевых этапов. Сначала мы проводим статистический анализ лингвистических признаков, влияющих на сложность текстов, на основе данных из бенчмарков IFEval, BBH, MMLU-Pro и GPQA. Затем применяем как традиционные, так и модифицированные формулы оценки удобочитаемости для анализа распределения оценок по текстам разной специфики (различные задачи и доменные области). Мы также изучаем корреляцию между оценками удобочитаемости / лингвистическими признаками и результатами языковых моделей. Наконец, мы исследуем возможность использования больших языковых моделей (LLM) для оценки удобочитаемости.

Результаты исследования показывают наличие корреляций между лингвистическими характеристиками и жанровыми особенностями задач или доменных областей. Для традиционных формул удобочитаемости (Flesch Grade, ARI, Lix, Dale-Chall, Lensear Write) характерен достаточно низкий разброс значений в рамках одного набора данных. Также мы предлагаем новый признак сложности текста, демонстрирующий более равномерное распределение на неоднородных данных. Использование LLM (в частности, GigaChat) для оценки удобочитаемости оказалось ограниченным: полученные оценки можно рассмотреть только как приблизительные, и, вероятно, модель не может быть использована без дообучения или использования специальных техник промптинга. Отметим, что оценки LLM на текстах, наиболее

близких к стилистике бытовой речи, продемонстрировали наименьшую корреляцию с традиционными формулами удобочитаемости.

ОЦЕНКА СВЯЗИ НАРРАТИВНОСТИ И ДОЛИ ПРОИЗВОДНЫХ СЛОВ В ХУДОЖЕСТВЕННЫХ И НАУЧНЫХ ТЕКСТАХ

Ю.А. Харлан

Томский государственный университет, Томск, Россия

В представленном исследовании дана оценка связи между такими показателями сложности текста, как нарративность и доля производных слов. Считается, что научные тексты характеризуются более низкими показателями нарративности, чем художественные, что может быть обусловлено не только увеличением долей определенных частеречных классов, но увеличением долей производных слов (девербативов и деадъективов).

В качестве материала исследования был использован корпус художественных текстов (рассказов и повестей), входящих в школьную программу для учащихся 10–11-х классов, а также корпус научных текстов – статей из журнала «Литература в школе». Общее количество художественных текстов – 38, научных – 46. Общее количество словоупотреблений в научном корпусе – 427 997, художественном – 121 828.

Нарративность рассчитывалась как отношение общего количества глаголов к общему количеству существительных. Для этого была выполнена морфологическая разметка с помощью программы TagAnt. При подсчете количества производных слов учитывались леммы вида *ция, *ние, *вие, *тие, *ист, *изм, *ура, *ище, *ство, *ость, *овка, *атор, *итор, *тель, *льный, *овать и др.

Анализ данных в обоих корпусах показал, что существительные составляют 22% от общего количество всех словоупотреблений в художественном корпусе, а глаголы – 18%, при этом 30% существительных являются производными. Процент существительных в научных текстах равен 30%, а процент глаголов – 12%, при этом практически половина существительных являются производными (47%). Средний показатель нарративности художественных текстов составляет 0,85, а научных – 0,39.

Для оценки связи нарративности и доли производных слов использовался корреляционный анализ, в результате которого было установлено, что нарративность обоих корпусов имеет относительно высокий показатель корреляции с долями производных слов (коэффициент корреляции Спирмена между показателями нарративности и долей производных слов в художественных текстах равен $-0,61$, показатель для корпуса научных текстов – $-0,70$). Также была обнаружена относительно высокая положительная корреляция между долей существительных и долей производных слов в обоих корпусах.

Таким образом, в ходе исследования было подтверждено, что научные тексты характеризуются более низкими показателями нарративности, чем художественные. Также подтвердилась гипотеза о том, что низкие показатели нарративности обусловлены увеличением долей производных слов.

THE IMPACT OF DIFFERENT TEXT PREPROCESSING PROCEDURES IN TEXT'S SENTIMENT TRAJECTORIES

A.D. Egorova¹, N.S. Nazarov¹

¹ HSE University, St. Petersburg, Russia

Sentiment analysis, a key component of Natural Language Processing (NLP), is traditionally used to classify texts or its components

based on its emotional tone. However, recent studies have demonstrated the potential of combining sentiment analysis with complexity measures such as entropy and semantic trajectories to gain deeper insights into text structure and meaning.

This research explores sentiment and semantic trajectories as effective tools for text analysis, particularly in evaluating literary translations and ranking books based on the extent to which the text is lively and chaotic and how much information is there. By applying entropy-complexity metrics to sentiment trajectories, we assess the degree of unpredictability and information dispersion within texts. In the present study probabilistic approach according to Shannon entropy and algorithmic approach of Kolmogorov (Kolmogorov complexity) will be considered. Sentiment analysis was employed in work to generate time series representing the emotional content of the text. These time series will subsequently be used to quantify textual complexity and entropy

One of the key findings of this study is that sentiment trajectories exhibit patterns of stochasticity and chaos, which can be quantified through mathematical techniques of Shannon entropy and Lyapunov exponents. These indicators enable a novel approach to evaluating translation quality by comparing the structural complexity of original texts and their translations. Our analysis of literary bestsellers, including works by Fyodor Dostoevsky and Stephen King, reveals significant variations in entropy-complexity measures depending on the translation method used - human translation, machine translation, or word-by-word automated translation.

After we carried out sentiment analysis and calculated measures of polarity and subjectivity, we found a time series from which we found entropy and normalized it. We performed this operation for one of Dostoevsky's bestselling books "Crime and Punishment" with and without preprocessing. We observed that without preprocessing, the entropy measure is lower, most likely due to the fact that stop-words that have the most pronounced emotiveness were removed.

This research contributes to the growing field of computational linguistics by demonstrating the applicability of mathematical models in sentiment analysis. The findings highlight the importance of integrating linguistic and mathematical approaches in NLP to enhance the accuracy of sentiment analysis. Future research will explore the transferability of these methods across different text genres and domains.

**READING MOTIVATION, READING SKILLS
AND PROBLEMATIC INTERNET USE IN ADOLESCENTS
IN GRADES 5–9: EVIDENCE FROM A PILOT STUDY
IN THREE RUSSIAN REGIONS***

M.S. Vlasov¹, O.A. Sychev¹

¹ *Altai State Pedagogical University, Biysk branch
named after V.M. Shukshin, Biysk, Russia*

The desire for online communication, excessive consumption of entertainment and information content from the Internet, acquiring a dominant position in the system of motives of adolescents, can displace other important motives, such as interest in studying or the desire to read literature. Based on the results of previous studies, we hypothesized that problematic Internet use (PIU) in schoolchildren is associated with reduced autonomous motivation for reading and increased controlled motivation and amotivation.

Autonomous motivation, associated with a sense of autonomous regulation, when the reasons for involvement in an activity are experienced as personal, not related to external control, includes intrinsic

* The study was supported by the Ministry of Education of the Russian Federation as part of the implementation of the state assignment for applied research “The relationship between problematic Internet use, reading motivation and reading skills of secondary school students in urban and rural schools” (state assignment No. 073-00014-24-07, No. PTNI 1023032200007-4-5.1.1; 6.2.6, state registration number of research 124120300005-4).

motivation and identified regulation. The latter is considered a productive type of extrinsic motivation. *Controlled motivation* includes extrinsic motivation and introjected regulation, reflecting the regulation of activity by introjected goals and values, perceived as somewhat alien, not one's own (for example, learned parental demands and expectations). In addition, along with autonomous and controlled motivation, the updated self-determination theory also distinguishes a state of *amotivation* as the absence of motivation for action, expressed either in the absence of actions as such, or in their formal implementation in the absence of their own motives, when people simply perform their duties.

The consequences of possible effects of PIU on motivation may affect a decrease in reading amount and reading skills.

Thus, our study aimed to analyze the relationship between PIU and reading motivation, reading amount and reading skills. The study involved 446 students in grades 5-9 from Biysk, Gorno-Altaysk, Perm, and rural areas of the Altai Territory and the Altai Republic.

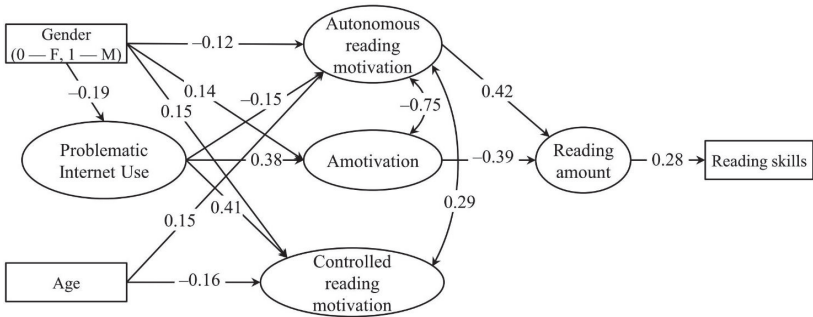


Figure 1. The tested structural model (all the coefficients shown are statistically significant at $p < 0.001$, insignificant paths and residuals are omitted for parsimony)

The study used the PISA-like reading literacy assessment, General Problematic Internet Use Scale-3 (GPIUS3) developed by

A.A. Gerasimova and A.B. Kholmogorova, SRQ-Reading Motivation, Reading amount Scale. The results of the correlation analysis showed the expected relationships between the indicators of PIU, reading motivation, reading amount, and reading skills.

Structural linear modeling was performed to test the hypothesis.

This model showed satisfactory fit to data: $\chi^2 = 821.32$, $df = 381$, $p < 0.001$, CFI = 0.912, TLI = 0.899, SRMR = 0.063, RMSEA = 0.051, 90% CI for RMSEA: 0.046-0.056, PCLOSE = 0.371, N = 446.

The model demonstrates that the direct effects of motivation and PIU on reading skills were not statistically significant. The direct effect of PIU on reading volume was also not significant. At the same time, the analysis of indirect effects showed that the effects of autonomous motivation (0.100; $p \leq 0.01$) and amotivation (-0.079; $p \leq 0.01$) on reading skills were significant. The total indirect negative effect of PIU was statistically significant for both reading skills (-0.044; $p \leq 0.01$) and reading amount (-0.198; $p \leq 0.001$). Thus, the model demonstrates that PIU is directly related to all indicators of reading motivation and indirectly through them to reading amount and reading skills.

**ЛЕКСИЧЕСКАЯ РЕПРЕЗЕНТАЦИЯ ПАНДЕМИИ
НОВОЙ КОРОНАВИРУСНОЙ ИНФЕКЦИИ:
МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА
ДЛЯ АНАЛИЗА СТРАТЕГИИ РИСК-КОММУНИКАЦИИ
ИНСТИТУТОВ ВЛАСТИ В ДЕЛОВОМ ДИСКУРСЕ**

Е.К. Плешкова

Сибирский государственный медицинский университет, Томск, Россия

Инфодемия – это бесконтрольно существующие потоки информации, характеризующейся наличием множества противоположных мнений, оценок, слухов, недостоверной информации.

Впервые об инфодемии как о социальном феномене весь мир заговорил во время пандемии новой коронавирусной инфекции.

В данном исследовании предполагается, что в потоках информации могут возникать смысловые пробелы, создающие риски для жизнедеятельности общества. Эффективное управление такими рисками в информационном пространстве возможно лишь при наличии грамотно выстроенной стратегии риск-коммуникации. Анализ разрывов в риск-коммуникации в условиях инфодемии и причин их возникновения может стать основой для разработки стратегий их предотвращения, а также для выработки механизмов взаимодействия социальных институтов и общества для решения этой проблемы.

В данной статье представлена лексическая репрезентация пандемии в деловом дискурсе. Сформирован датасет для выявления репрезентантов пандемии в деловом дискурсе. Были отобраны тексты документов (распоряжения, постановления, клинические рекомендации и т.д.), размещенные в открытом банке документов на сайте Министерства здравоохранения Российской Федерации и Роспотребнадзора РФ, за период с 01.03.2020 по 01.09.2020. Объем текстов составил 47 292 слова. На этапе предобработки текстов были удалены гиперссылки, встречавшиеся в документах, знаки препинания и цифры. Была проведена токенизация корпуса текстов, на следующем этапе из корпуса были исключены стоп-слова, выполнена лемматизация с использованием библиотеки Snowball. С помощью функции `topfeatures` («Quanteda») выявлены наиболее частотные лексические единицы. На основе лемматизированных токенов создана матрица `datafeature matrix` (77 токенов), в которой остались только те токены, которые встречались более 100 раз. Далее с использованием функции `textplot_wordcloud` было создано «облако слов», в котором собраны ключевые лексемы, репрезентирующие пандемию в деловом дискурсе (рис. 1).



Рис. 1. Диаграмма «облако слов», включающая лексические единицы, встречавшиеся более 100 раз в текстах делового дискурса в период с 01.03 по 01.09.2020 г.

Диаграмма «облако слов» визуализировала лексические единицы, наиболее распространенные в текстах делового дискурса, опубликованных в ответ на развитие пандемии, направленных на контроль и сдерживание пандемии. Как следствие, данные лексемы визуализируют темы, наиболее важные для российских институтов власти в период пандемии, посредством которых они выстраивали свои стратегии риск-коммуникации.

АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ АСПЕКТОВ В РАМКАХ ЗАДАЧИ ASPECT-BASED SENTIMENT ANALYSIS НА МАТЕРИАЛЕ ОТЗЫВОВ О ХУДОЖЕСТВЕННЫХ ПРОИЗВЕДЕНИЯХ С САЙТА LITRES

Е.И. Зима¹, Я.С. Холодов¹

¹ Томский государственный университет, Томск, Россия

В современном мире цифрового маркетинга основным источником знаний для продвижения продукта или услуги служат открытые данные, получаемые из Интернет-сервисов. Одним из наиболее распространённых видов опосредованной коммуникации пользователей с бизнесом является Интернет-жанр отзыва. В зависимости от предоставляемой услуги содержательная часть отзыва может быть связана как с качеством товара (пр. наличие брака), так и с выражением специфических впечатлений о приобретенном опыте. В частности, к сфере читательского опыта можно отнести отзыв о художественном произведении, где мнение может отличаться в зависимости от определенного аспекта, пр. сюжет, персонажи, стиль написания, качество издания и т.д.

Проблема автоматического определения тональности составляющих текста уже получила широкое внимание со стороны исследователей, и в англоязычной литературе была обозначена как aspect-based sentiment analysis. Однако особая сложность такого подхода заключается в выделении конечного списка оцениваемых с точки зрения тональности сегментов отзывов, которые могут быть специфичными для разных типов текстов. Уникальность данной проблемы заключается в том, что, в отличие от Интернет-отзывов на товары с маркетплейсов, где критерии оценки часто

стандартизированы (например, «цена», «качество», «качество доставки»), в отзывах на художественные произведения аспекты могут зависеть от жанра книги, авторского стиля и индивидуальных читательских ожиданий.

В качестве решения поставленной проблемы мы предлагаем метод тематического моделирования (Topic Modeling) как один из подходов в рамках задачи кластеризации данных. В частности, нами было проведено сравнение моделей Latent Dirichlet Allocation (LDA) и BERTopic, оптимизированных с учетом специфики русскоязычных текстов и стилистическими особенностями книжных отзывов. Принимая гипотезу о том, что объекты оценки могут совпадать с тематическим распределением в отзывах, мы выявили первичный список аспектов, относящихся к содержанию художественного произведения: герой, сюжет, сцены, сеттинг, читательский опыт, любовная линия, тональность повествования.

Учитывая преимущества и недостатки рассмотренных методов тематического моделирования, следующим этапом исследования нам видится применение модели semi-supervised LDA с учетом уже выделенных тем для анализа их лексического наполнения и выделения более специфических аспектов в отзывах.

ЯЗЫКОВЫЕ МОДЕЛИ В ГЕНЕРАЦИИ И ОБРАБОТКЕ ТЕКСТОВ

ПРЕДСКАЗАНИЕ УСПЕШНОСТИ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ ПО ЭМОЦИОНАЛЬНОЙ КРИВОЙ СЮЖЕТА

М.А. Кутейникова¹, А.Л. Касьяненко²

¹ *Томский государственный университет, Томск, Россия*

² *Издательство «Эксмо», Москва, Россия*

В этой работе представлена система автоматизированного рецензирования художественных рукописей, поступающих в издательство, основанная на анализе эмоционального восприятия текста, и сфокусированная, в первую очередь, на выделение этапов сюжета. Разработка подобных методов особенно актуальна в условиях возрастающего количества подаваемых рукописей и необходимости быстрой предварительной оценки их литературной ценности. Для достижения цели текст разбивается на предложения, и каждое предложение анализируется с помощью модели DeepPavlov для определения его эмоционального окраса. На основе полученных данных строится эмоциональная кривая произведения, график, отражающий динамику эмоций текста, построенный на основе анализа эмоциональной тональности предложений. Анализ этой кривой позволяет выявить ключевые точки повествования, такие как завязка, кульминация и развязка, а также определить, насколько гармонично распределены эмоциональные всплески.

Эмоциональная кривая играет одну из ключевых ролей в предсказании успеха произведения. Анализ показал, что определённые паттерны кривой коррелируют с вероятностью публикации книги и её последующим коммерческим успехом. В исследовании использован размеченный датасет рукописей, предоставленный издательством, где произведения классифицируются как успешные (бестселлеры), опубликованные, но не получившие широкого распространения, и отклонённые. Построенная модель машинного обучения классифицирует книги по категориям успешности на основе анализа их эмоциональных кривых.

Результаты показали, что использование этого подхода позволяет с хорошей точностью прогнозировать успех книги ещё до её публикации. Это может значительно сократить время на рецензирование и помочь издательствам принимать более обоснованные решения о публикации. Методология также может быть применена для сценарного анализа, создания рекомендаций по улучшению рукописи и выявления слабых мест в сюжете. Таким образом, предложенный подход представляет собой перспективный инструмент для автоматизированного анализа литературных произведений, объединяя машинное обучение и анализ эмоций для объективной оценки художественного качества текста. Он открывает возможности не только для ускорения работы редакторов, но и для более точного выявления книг с высоким потенциалом. Для более точного предсказания успешности книги, на взгляд авторов, необходимо добавить дополнительные методы автоматизированного анализа, такие как морфологический и лексический анализ.

Авторы выражают благодарность Томскому государственному университету за предоставленную возможность провести вычисления на суперкомпьютере «СКИФ Cyberia».

АВТОМАТИЧЕСКОЕ АННОТИРОВАНИЕ ОШИБОК В УСТНОЙ РЕЧИ НА АНГЛИЙСКОМ ЯЗЫКЕ С ПРИМЕНЕНИЕМ LLM

Ю.С. Виноградова¹, И.В. Ильченко¹, П.С. Ширяева¹

¹ Высшая школа экономики, Санкт-Петербург, Россия

Автоматическое аннотирование устной речи продолжает быть актуальной задачей в области лингвистических исследований и технологий обработки естественного языка. Существующие решения по выявлению ошибок в иноязычной речи преимущественно представлены коммерческими продуктами с ограниченным доступом, что затрудняет их использование в образовательных целях.

Развитие больших языковых моделей (LLM) значительно расширило возможности автоматизированного аннотирования, обеспечивая более точный анализ речи и адаптивную обратную связь. Разметка ошибок в иноязычных текстах представляет собой сложную задачу, требующую учета контекстуальных факторов, уровня языковой компетенции и типичных ошибок, характерных для носителей определенного языка. Использование методов машинного обучения для решения данной проблемы требует значительных вычислительных ресурсов. В этом контексте корпуса ошибок с ручной разметкой выступают в качестве ценных данных для последующего автоматического аннотирования, что повышает эффективность обработки устной речи и минимизирует необходимость трудозатратной ручной разметки.

В рамках настоящего исследования разработан инструмент для выявления ошибок в устной речи на иностранном языке. Система принимает на вход расшифровку английской речи и предоставляет перечень ошибок, классифицированных по уровням

языка. В основе решения лежит GPT-ассистент, который представляет собой систему, основанную на большой языковой модели (LLM) и адаптированную под конкретную задачу. Наш ассистент предназначен для обработки расшифровок устной речи и автоматического выявления ошибок. Обучение модели проводилось на корпусе устной речи русскоязычных студентов, изучающих профессиональный английский язык в сфере информационных технологий. Корпус был создан и аннотирован на основе материалов онлайн-уроков и отражает распространенные ошибки данной группы обучающихся. Настройка модели осуществлялась с применением методов промптинга.

На текущем этапе проводится интеграция системы автоматизированной разметки ошибок с платформой Telegram для создания удобного пользовательского интерфейса в виде чат-бота. Пользователь может прислать боту голосовое сообщение на английском языке, которое преобразуется в текст с помощью Whisper. Расшифрованное сообщение отправляется GPT-ассистенту, который выявляет основные ошибки и возвращает пользователю их список с классификацией по типам. Планируемый инструмент ориентирован на широкий круг пользователей, а интерактивный формат и мобильная доступность обеспечат удобство самостоятельного обучения и совершенствования языковых навыков. В дальнейшем планируется внедрение модуля статистического анализа, позволяющего пользователям отслеживать динамику собственных ошибок и их исправлений.

ДИСТИЛЛЯЦИЯ НА УРОВНЕ ТЕКСТОВ: КАК СОЗДАТЬ МАЛЕНЬКИЙ, НО ЭФФЕКТИВНЫЙ ЭНКОДЕР ПРЕДЛОЖЕНИЙ

Б.Т. Малашенко

*Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики (ИТМО),
Санкт-Петербург, Россия*

Кодирование текстов остаётся одной из ключевых задач в области обработки естественного языка. Векторные представления текстов – эмбединги – широко используются для классификации, кластеризации, поиска, суммаризации и других задач. Недавние исследования показали, что удаление каузальной маски в больших языковых моделях (БЯМ) и их дообучение под задачу кодирования позволяет создавать энкодеры с миллиардами параметров. Однако для задач векторного поиска требуется высокая пропускная способность модели и низкая размерность латентного пространства, особенно при работе с большими базами данных, содержащими сотни миллиардов скрытых представлений.

Одним из решений является дистилляция на уровне текстовых представлений, при которой модель-ученик заимствует знания от более крупной модели-учителя. Это позволяет сохранить высокое качество при уменьшении размера модели. Возможно использование нескольких учителей для дистилляции в одну модель-ученика. Важно, чтобы в процессе дистилляции не нарушались представления модели-ученика, если она уже прошла фазу предобучения. Для этого перед дистилляцией необходимо спроецировать латентное пространство учителя на пространство ученика, минимизируя различия в начальной инициализации векторов.

Результаты экспериментов показали, что дистилляция на уровне текстовых представлений позволяет создать качественную и быструю модель. В качестве модели ученика использована архитектура, аналогичная bert_uncased_L-2_H-128_A-2, а модель rubert-tiny использована как бейзлайн. После дистилляции удалось превзойти бейзлайн по агрегации метрик на бенчмарке RuMTEB.

ОПТИМИЗАЦИЯ FINE-TUNING'A: СООТНОШЕНИЕ ПАРАМЕТРОВ МОДЕЛИ И ТОКЕНОВ (НА ПРИМЕРЕ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ И ЛИДОВ)

Ф.Ф. Шамигов

Томский государственный университет, Томск, Россия

Исследование посвящено проблеме оптимизации дообучения больших языковых моделей (БЯМ) для автоматической генерации заголовков и лидов новостных статей на материале русского языка.

Актуальность работы обуславливается значимостью проблемы оптимизации обучения БЯМ. Мы стремимся внести вклад в исследования по оптимизации обучения моделей, в фокусе которых балансирование производительности и практичности.

Также эта работа делает вклад в развитие систем автоматической генерации новостных заголовков и лидов – очень ценного и привлекательного инструмента для СМИ.

Цель – установить, есть ли разница в качестве генерируемого заголовка и лида между «мощной» моделью ruGPT-3 Medium, дообученной на меньшем количестве данных, и «слабой» моделью ruGPT-3 Small, дообученной на большем количестве данных.

В настоящем исследовании применялись **методы** предобработки данных (отбор рубрик и количества статей, удаление ненужных колонок, разделение на датасеты по рубрикам, разделение на обучающую и тестовую выборки, перевод в формат json), метод fine-tuning'а большой языковой модели (дообучение на задаче суммаризации), методы оценки качества работы нейронных сетей (ROUGE, BLEU, chrF, и BERTScore). Использовался язык программирования Python в сервисе Kaggle.

Результаты анализа по метрикам для **заголовков** представлены в табл. 1.

Т а б л и ц а 1

RuGPT-3 Medium, 350 млн параметров, 5 900 статей				
Рубрика	ROGUE	BLEU	ChrF	BERTScore
Наука	17.6	9.4	33.1	86.6
Спорт	19.9	10	34.9	86.9
Россия	24.2	12.7	40.5	88.3
Экономика	24.7	13.8	39.4	88.7
Инт. и СМИ	24	13.6	37.7	88
RuGPT-3 Small, 125 млн параметров, 1 5000 статей				
Рубрика	ROGUE	BLEU	ChrF	BERTScore
Наука	23.8	13.3	37.2	87.8
Спорт	32.8	19.8	44.2	89.5
Россия	22.8	11.8	38.4	87.9
Экономика	22.6	12.4	36.6	88.2
Инт. и СМИ	21.4	11.8	36.7	87.5

Результаты анализа по метрикам для **лидов** представлены в табл. 2.

Таблица 2

RuGPT-3 Medium, 350 млн параметров, 4 800 статей				
Рубрика	ROGUE	BLEU	ChrF	BERTScore
Экономика	13.6	5.5	29	86
Происшествия	15.8	6.9	30	87
Спорт	17.2	8.3	30.6	87.5
Культура	10.2	3.7	21.5	84.7
Мир	16.9	7.3	30.6	87.2
RuGPT-3 Small, 125 млн параметров, 12 000 статей				
Рубрика	ROGUE	BLEU	ChrF	BERTScore
Экономика	12.3	4.9	26	85.2
Происшествия	14.3	6	29	86.5
Спорт	14.9	7.1	28	86.7
Культура	8.4	3	19.4	83.9
Мир	15.3	6.6	28.5	86.6

В результате для заголовков в 3 из 5 рубрик «сильная» модель ruGPT-3 Medium (350 млн параметров), дообученная на меньшем количестве данных (4 800–5 900 статей), показала лучший результат по всем метрикам в сравнении со «слабой» моделью ruGPT-3 Small (125 млн параметров), дообученной на большем количестве данных (12 000–15 000 статей); а для лидов – «сильная» модель показала лучший результат по всем метрикам в каждой рубрике.

АНАЛИЗ ПРИЁМОВ АРГУМЕНТАЦИИ В ТЕКСТАХ НАУЧНОЙ КОММУНИКАЦИИ

И.С. Пименов

Институт систем информатики им. А.П. Ершова, Новосибирск, Россия

Исследование обращено к проблеме автоматического выявления приёмов аргументации из аргументационных графов для тек-

стов трёх жанров: научных и научно-популярных статей, новостях науки. Под приёмами аргументации понимаются повторяющиеся между разными текстами структуры из типовых моделей рассуждения, реализуемых в связанных аргументах.

В эксперименте использован корпус из 98 текстов с метками жанров: 28 новостей науки (PoiskNews), 20 статей Nabr, 50 научных статей двух тематических областей (по 25 из лингвистики и информационных технологий в открытом доступе в библиотеке «КиберЛенинка»). Корпус содержит 5042 утверждения и 4328 аргументов. Разметка аргументации в текстах корпуса осуществлена пятью экспертами согласно стандарту Argument Interchange Format и компендиуму моделей рассуждения Дугласа Уолтона. В результате разметки аргументация в каждом тексте представляется в виде связного ориентированного графа.

Задача выявления приёмов аргументации сводится к извлечению изоморфных подграфов из преобразованных аргументационных графов, содержащих только вершины с моделями рассуждений. Применены методы частотного анализа подграфов с проверкой на изоморфизм через реализацию алгоритма Корделлы VF2. Результатом является частотный спектр приёмов аргументации (содержит подграфы, соответствующие приёмам, и их жанровую, текстовую, абсолютную частоту). Из спектра исключены вложенные приёмы, функционирующие только в составе более крупных. Для оценки, насколько полно приёмы характеризуют тексты, проведена кластеризация текстов по приёмам методами Ward и K-means.

Из анализа частотного спектра установлено, что межжанровых приёмов значительно меньше, чем специфичных одному жанру (88 и 767: 579 в научных статьях, 132 в Nabr, 56 в новостях). Наиболее близка аргументация в научных и научно-популярных статьях, наименее – в научно-популярных текстах и научных новостях. Научные статьи характеризуются наибольшим

разнообразием и частотностью составных приёмов, научно-популярные отличаются активной полемической аргументацией, доказательства в научных новостях ведутся компактными приёмами с низким разнообразием.

Показано, что кластеризация на основе приёмов надёжно разграничивает тексты по жанрам и тематическим особенностям (три кластера из пяти образованы исключительно текстами одного жанра, два других – преимущественно одного, к которому принадлежат 97 и 76% текстов кластера), позволяет выявлять тексты со схожей аргументацией.

Результаты анализа приёмов аргументации значимы для исследования более сложных стратегий аргументации, характеризующих целостные тексты, могут использоваться для прогнозирования убедительности текстов, генерации аргументации.

ПРОМПТ-ИНЖИНИРИНГ, РАЗРАБОТКА ЧАТ-БОТОВ И ПРОЧИЕ ПРИКЛАДНЫЕ NLP-ЗАДАЧИ

ПРОТОТИП ЧАТ-БОТА ДЛЯ ВЛАДИМИРО-СУЗДАЛЬСКОГО МУЗЕЯ-ЗАПОВЕДНИКА

Л.Д. Тельнина

Высшая школа экономики, Санкт-Петербург, Россия

В условиях цифровизации культурных учреждений автоматический сбор и анализ обратной связи от посетителей приобретает ключевое значение для совершенствования музейной коммуникации. Автоматизация данного процесса позволяет как сократить временные затраты, так и систематизировать обработку данных, что способствует скорейшему обнаружению точек развития. В рамках проектной работы для Владимиро-Суздальского музея-заповедника была разработана прототипная версия чат-бота, предназначенного для опроса посетителей музея и последующей записи данных в таблицу.

При работе над ботом была спроектирована схема сценария взаимодействия человек-робот, которая отображает как «золотой путь», подразумевающий прохождение сценария до запланированного конца и достижения цели общения, так и вероятные отклонения. Кроме этого, была подготовлена персона бота и его стиль общения. Что касается функциональности, прототип, разработанный при помощи инструментов Just AI Conversational

Platform (JAICP), может задавать вопросы; фиксировать текстовые и кнопочные ответы на них и через интеграцию сохранять в Google Sheets в нужные столбцы; сохраняя контекст, отрабатывать нераспознанные запросы на разных этапах диалога; завершать диалог по превышении времени ожидания, а также адаптировать ход диалога в зависимости от знакомства с клиентом. Для удобства взаимодействия и в связи с большой аудиторией чат-бот размещен в Telegram. Модуль Natural Language Understanding (NLU) реализован с использованием регулярных выражений (RegEx).

Предполагается, что чат-бот, подготовленный на основе разработанного прототипа, позволит ускорить процесс сбора данных и масштабировать их объем. Таким образом, музей сможет располагать достаточным количеством структурированной информации, с помощью которой можно будет выявить ключевые тенденции и определить направления для развития в работе музея.

РАЗРАБОТКА ЧАТ-БОТОВ ДЛЯ ПЕРСОНАЛИЗИРОВАННОГО ИЗУЧЕНИЯ ИНОСТРАННЫХ ЯЗЫКОВ: БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ

А.П. Золин

Высшая школа экономики, Санкт-Петербург, Россия

Обработка естественного языка (NLP) играет важную роль в современном мире, находя применение в различных областях, включая сферу образования. Одним из наиболее перспективных направлений в данной области является использование чат-ботов для изучения иностранных языков. Современные чат-боты явля-

ются очень удобной платформой для реализации персонализированного обучения, поскольку их функциональные возможности и обширный инструментарий позволяют обеспечить гибкую настройку учебного процесса.

Для выявления современных тенденций в этой области был проведен библиометрический анализ. Поиск источников был осуществлен в базе данных Dimensions, благодаря которой удалось найти 13618 релевантных научных публикаций по следующим ключевым словам: «personalization», «chatbot» и «language learning». Затем с помощью инструмента визуализации VOSviewer была построена библиометрическая карта, на которой явно выделяются 5 основных кластеров, отражающих ключевые направления исследований в данной области.

Первый кластер олицетворяет сферу образования, а также подчеркивает тенденцию к активному внедрению новых технологий (в т.ч. искусственного интеллекта) в учебный процесс.

Второй кластер посвящен пользовательскому взаимодействию и восприятию внедрения новых технологий. Также он связан с методами оценки эффективности чат-ботов, которые стоит учитывать для измерения степени персонализации и влияния чат-ботов на обучение.

Третий кластер отражает психологические аспекты взаимодействия пользователей с чат-ботами, такие как уровень стресса, вовлеченность и прочие индивидуальные особенности. При разработке персонализированных решений очень важно учитывать различные когнитивные и эмоциональные факторы.

Четвертый кластер посвящен технической составляющей современных чат-ботов, а также новых методах, которые активно применяются при их проектировании.

Пятый кластер демонстрирует коммерческое применение чат-ботов, включая автоматизацию обслуживания клиентов и интеграцию в коммерческие модели.

Результаты исследования позволяют определить степень работанности проблемы, а также выявить основные тренды для усовершенствования дальнейших разработок в сфере персонализированного изучения иностранных языков с использованием чат-ботов.

РАЗРАБОТКА ЧАТ-БОТА ДЛЯ ЭРМИТАЖА: АВТОМАТИЗАЦИЯ ВЗАИМОДЕЙСТВИЯ С ПОСЕТИТЕЛЯМИ МУЗЕЯ

М.О. Сергеева¹, Е.А. Сытикова¹

¹ Высшая школа экономики, Санкт-Петербург, Россия

Современные музеи стремятся к улучшению взаимодействия с посетителями, используя цифровые технологии. В данной работе представлено описание чат-бота для Государственного Эрмитажа, разработанный для автоматизации ответов на часто задаваемые вопросы и улучшения пользовательского опыта.

Основная задача проекта заключается в предоставлении пользователю актуальной информации о музее, его экспонатах, услугах и расписании по запросу пользователя. В настоящий момент база знаний бота содержит часто задаваемые вопросы посетителей музея, собранные колл-центром и сотрудниками отдела информационного сопровождения Эрмитажа, данные об истории музея, а также текущая информация о расписании его работы.

Генерация ответов на запросы пользователя происходит с помощью технологии ассистента от Яндекса, в основе которой лежит метод Retrieval-Augmented Generation (RAG). При помощи RAG ответы генерируются на основе базы знаний о музее и не

позволяет ассистенту использовать данные из открытых источников. Таким образом, пользователь, обращаясь к боту, получает достоверную информацию.

Для настройки стиля общения бота и обеспечения соответствия ответов стилю коммуникации (tone-of-voice) музея нами был также применен метод Low-Rank Adaptation (LoRA), который позволил дообучить языковую модель.

Благодаря LoRA мы смогли адаптировать ответы чат-бота таким образом, чтобы они были информативными и соответствовали статусу музея, но при этом оставались понятными и удобными для всех категорий посетителей.

Для повышения качества ответов реализованы механизмы контроля достоверности информации и ведения логов взаимодействия, которые содержат в себе информацию о совершенных пользователем действиях в рамках одной сессии, заданных вопросах и полученных ответах, а также времени, потраченном на ответы. Эти данные позволят анализировать пользовательские сценарии, выявлять наиболее частые запросы и улучшать качество ответов системы.

В дальнейшем данные из логов будут использоваться для анализа и оптимизации работы чат-бота, а также для выявления возможных проблем в его функционировании.

Проведенное альфа-тестирование на ограниченной группе пользователей показало высокий уровень удобства и информативности чат-бота. Пользователи отметили его способность давать точные и развернутые ответы, а также простоту навигации по музейной информации.

В ходе дальнейших этапов разработки запланировано проведение бета-тестирования с целью анализа реальных пользовательских запросов и дальнейшего совершенствования диалоговой модели.

Финальная версия чат-бота будет доступна в официальном Telegram-канале Эрмитажа, обеспечивая круглосуточную поддержку посетителей музея.

Ожидается, что внедрение данной технологии повысит доступность информации, упростит процесс взаимодействия с музеем и сделает посещение Эрмитажа более комфортным и информативным.

НЕЙРОСЕТЕВОЕ OCR-РАСПОЗНАВАНИЕ ИСТОРИЧЕСКИХ ТЕКСТОВ (НА МАТЕРИАЛЕ ГАЗЕТЫ «ТОМСКИЙ ВЕСТНИК»)

Д.В. Ващенко¹, А.В. Куценко¹

¹ Томский государственный университет, Томск, Россия

В докладе рассматривается нейросетевой подход к распознаванию исторических текстов с использованием технологий компьютерного зрения. Материалом для распознавания послужила газета «Томский вестник», отражающая общественно-политическую жизнь города Томска начала XX в. Газета выходила с 1912 по 1915 г., поэтому задача стояла распознать русский текст в дореформенной орфографии. Существующее программное обеспечение для оптического распознавания символов (OCR) не заточено на распознавание русской дореформенной орфографии (дореволюционной орфографии). Таким образом, исследование альтернативных подходов к OCR остаётся актуальным.

В исследовании применяются две нейросетевые модели, предназначенные для решения задач распознавания текстовых документов: Easy OCR и Tesseract OCR Engine. Модели находятся в открытом доступе. Их можно использовать напрямую или с помощью API для извлечения печатного текста из изображений.

Для формирования текстовой коллекции был взят выпуск «Томского вестника» № 36 от 7 октября 1912 г. В ходе предобработки изображений страниц было выделено 102 самостоятельных текстовых сегмента, которые итеративно подавались в модель для распознавания. Итоговый датасет включает более 11 тыс. слов и более 81 тыс. символов соответственно.

Сравнительный анализ точности моделей проводился на основе метрик WER (Word Error Rate) и CER (Character Error Rate), которые измеряют расхождения между распознанным и эталонным текстами. Для подсчёта метрик был подготовлен эталонный датасет с исходным текстом газетных статей.

Исследование показало, что модель Tesseract демонстрирует лучшие результаты, с показателями WER 0.65 и CER 0.24, в сравнении с EasyOCR, имеющей показатели WER 0.89 и CER 0.41. Несмотря на наличие ошибок, большая часть символов и слов распознается верно.

Результаты исследования позволяют определить, какая из моделей OCR лучше справляется с распознаванием дореволюционной орфографии, и могут быть полезны для автоматизации обработки исторических документов. Однако для достижения более высокой точности распознавания требуется дополнительная посткоррекция результатов распознавания.

ИССЛЕДОВАНИЕ И ВЫБОР МОДЕЛИ ФИЛЬТРАЦИИ ТОКСИЧНЫХ СООБЩЕНИЙ ДЛЯ ЦИФРОВОГО АССИСТЕНТА ОБРАЗОВАТЕЛЬНОЙ ПЛАТФОРМЫ

Е.А. Андриевская¹, В.Е. Фомин¹, Д.А. Стрыгин¹

¹ «Образовариум», Москва, Россия

Цель исследования – разработка инструмента выбора оптимальной модели или ансамбля моделей для фильтрации токсичных сообщений в голосовом ассистенте образовательной платформы.

Задача сводилась к бинарной классификации: определению токсичности текста (агрессивный/нейтральный).

Для решения задачи были рассмотрены предобученные языковые модели, специализирующиеся на классификации токсичности русскоязычных текстов, включая RuBERT, XLM-RoBERTa и Detoxify.

В исследовании проанализированы модели с использованием различных порогов классификации, а также комбинированные подходы на основе логики И/ИЛИ.

Оценка эффективности моделей проводилась с использованием следующих метрик: точность (Precision), полнота (Recall) и F1-мера, где Precision позволяла контролировать уровень ложных срабатываний, Recall – минимизировать пропуск токсичных сообщений, а F1-мера использовалась для балансировки обоих показателей.

Для валидации применялись датасеты, имитирующие реальные пользовательские запросы к голосовым ассистентам, включая тексты с эмоциональной окраской.

Результаты показали, что выбор модели и порога классификации существенно влияет на эффективность фильтрации.

Использование ансамблей позволяет гибко настраивать баланс между Precision и Recall, что особенно важно для образовательных платформ.

Лучшие результаты продемонстрировали модели, дообученные на данных из форумов и социальных сетей, а также компактные решения, оптимизированные для работы в реальном времени.

Предложенный инструмент представляет собой программный модуль, интегрируемый в чат онлайн-ассистента, реализующий методику отбора и комбинирования моделей фильтрации токсичных сообщений в зависимости от выбранных метрик и порогов.

Разработанный подход позволяет адаптировать фильтрацию под специфику разных образовательных платформ и сценариев взаимодействия.

Перспективы включают дообучение моделей на специализированных образовательных данных и развитие ансамблевых методов для повышения точности фильтрации.

ПРОМПТ-ИНЖИНИРИНГ И СОВЕРШЕНСТВОВАНИЕ ВЗАИМОДЕЙСТВИЯ С ИНСТРУМЕНТАМИ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В.М. Бирюков

Томский государственный университет, Томск, Россия

Проблема, рассматриваемая в исследовании, заключается в ограниченной эффективности взаимодействия с генеративными моделями ИИ при использовании стандартных обобщённых запросов.

Для решения этой задачи требуется разработка и оптимизация промптов, которые способны значительно повысить точность выполнения задач и снизить количество ошибок (галлюцинаций). Вопрос, на который направлено исследование, состоит в том, как структурированные и контекстуально обогащённые запросы могут повлиять на улучшение качества работы моделей ИИ.

Для решения задачи был проведён экспериментальный анализ, включающий 18 современных моделей ИИ (OpenAI, Google Gemini, DeepSeek и Anthropic).

Основным методом исследования стал сравнительный анализ эффективности работы моделей с использованием базовых и детализированных промптов.

Метрики, использованные для оценки, включали точность выполнения задач, уровень ошибок (галлюцинаций) и время отклика.

Промпты были разработаны с использованием фреймворка CLEAR (Краткость, Логичность, Ясность, Адаптивность, Рефлексивность), что позволило стандартизировать входные данные.

Результаты исследования показали, что при использовании обогащённых промптов точность выполнения задач увеличилась на 35–50%, а количество ошибок (галлюцинаций) снижалось на 60%.

Например, точность медицинской диагностики с применением структурированных промптов повысилась с 68 до 93%. Также было выявлено, что использование многоступенчатых логических алгоритмов и контекстных инструкций способствовало повышению производительности и уменьшению времени отклика на 20–30%.

В дальнейшем предполагается развитие универсальных методов генерации промптов, а также создание адаптивных инструментов для динамической настройки запросов в зависимости от обновлений моделей ИИ.

Перспективы практического внедрения включают улучшение точности и надёжности ИИ в различных областях, таких как медицина, программирование и креативные индустрии.

СОДЕРЖАНИЕ

Предисловие (З.И. Резанова)	3
--	---

КОРПУСА И БАЗЫ ДАННЫХ ДЛЯ ИССЛЕДОВАТЕЛЬСКИХ ЗАДАЧ

<i>Биденко Э.А.</i> РАЗРАБОТКА СЛОВАРЯ-ТЕЗАУРУСА МИСТИКО-АСКЕТИЧЕСКОЙ ТРАДИЦИИ ИСИХАЗМА	6
<i>Зоркальцев А.В.</i> СОСТАВЛЕНИЕ КОРПУСА РИТОРИЧЕСКИХ СТРУКТУР АНГЛОЯЗЫЧНОЙ И РУССКОЯЗЫЧНОЙ ПЕСЕННОЙ ПОЭЗИИ И ЕГО ВОЗМОЖНЫЕ ПРИМЕНЕНИЯ	7
<i>Понкратова Е.М.</i> КОРПУСНЫЕ ТЕХНОЛОГИИ В ИЗУЧЕНИИ ТВОРЧЕСТВА Ф.М. ДОСТОЕВСКОГО	9
<i>Сушков Л.А., Трунцева Е.П.</i> СОЗДАНИЕ КОРПУСА DOSPENA ДЛЯ РАСПОЗНАВАНИЯ НАТЕ SPEECH НА МАТЕРИАЛЕ ЧАТА ВИДЕОИГРЫ DOTA 2	11
<i>Чекмез У.Э.</i> СОЗДАНИЕ КОРПУСА ЗВУЧАЩЕЙ РЕЧИ ДЛЯ ЛИНГВИСТИЧЕСКИХ И МЕЖДИСЦИПЛИНАРНЫХ ИССЛЕДОВАНИЙ	12
<i>Андреева М.И., Солнышкина М.И., Данилов А.В.</i> АНАЛИТИКА УЧЕБНЫХ ТЕКСТОВ ДЛЯ ИЗУЧЕНИЯ РУССКОГО ЯЗЫКА КАК ИНОСТРАННОГО: ФУНКЦИОНАЛ ПЛАТФОРМЫ RULINGVA	13
<i>Будянская Е.М., Бузанов А.О., Жорник Д.О., Пихтин А.А.</i> ОТ ПОЛЕВЫХ ДАННЫХ К АВТОМАТИЧЕСКОМУ АНАЛИЗУ: РАЗРАБОТКА МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА ДЛЯ НАУКАНСКОГО ЯЗЫКА	16
<i>Брок М.С.</i> СОВРЕМЕННЫЕ МЕТОДЫ ИСПОЛЬЗОВАНИЯ РАЗМЕТКИ ИМЕНОВАННЫХ СУЩНОСТЕЙ НА ПРИМЕРЕ РАЗМЕТКИ КОРПУСА ДЕТСКОЙ ЛИТЕРАТУРЫ ИЗ КАТАЛОГА БИБЛИОТЕКИ МАРТЬЯНОВА	18

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ ДЛЯ СОЦИАЛЬНО ЗНАЧИМЫХ ЗАДАЧ

<i>Мартынова Е.В., Куприянов Р.В., Солнышкина М.И.</i> СЛОЖНОСТЬ ТЕКСТА КАК МЕЖДИСЦИПЛИНАРНАЯ КАТЕГОРИЯ	21
<i>Засыпкин А.С., Бручес Е.П.</i> ПОДХОДЫ К ОЦЕНКЕ СЛОЖНОСТИ ТЕКСТОВ, ПРЕДНАЗНАЧЕННЫХ ДЛЯ ТЕСТИРОВАНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ	23
<i>Харлан Ю.А.</i> ОЦЕНКА СВЯЗИ НАРРАТИВНОСТИ И ДОЛИ ПРОИЗВОДНЫХ СЛОВ В ХУДОЖЕСТВЕННЫХ И НАУЧНЫХ ТЕКСТАХ	25
<i>Egorova A.D., Nazarov N.S.</i> THE IMPACT OF DIFFERENT TEXT PREPROCESSING PROCEDURES IN TEXT'S SENTIMENT TRAJECTORIES	26
<i>Vlasov M.S., Sychev O.A.</i> EADING MOTIVATION, READING SKILLS AND PROBLEMATIC INTERNET USE IN ADOLESCENTS IN GRADES 5–9: EVIDENCE FROM A PILOT STUDY IN THREE RUSSIAN REGIONS	28
<i>Плешкова Е.К.</i> ЛЕКСИЧЕСКАЯ РЕПРЕЗЕНТАЦИЯ ПАНДЕМИИ НОВОЙ КОРОНАВИРУСНОЙ ИНФЕКЦИИ: МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА ДЛЯ АНАЛИЗА СТРАТЕГИИ РИСК-КОММУНИКАЦИИ ИНСТИТУТОВ ВЛАСТИ В ДЕЛОВОМ ДИСКУРСЕ	30
<i>Зима Е.И., Холодов Я.С.</i> АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ АСПЕКТОВ В РАМКАХ ЗАДАЧИ ASPECT-BASED SENTIMENT ANALYSIS НА МАТЕРИАЛЕ ОТЗЫВОВ О ХУДОЖЕСТВЕННЫХ ПРОИЗВЕДЕНИЯХ С САЙТА LITRES	33

ЯЗЫКОВЫЕ МОДЕЛИ В ГЕНЕРАЦИИ И ОБРАБОТКЕ ТЕКСТОВ

<i>Кутейникова М.А., Касьяненко А.Л.</i> ПРЕДСКАЗАНИЕ УСПЕШНОСТИ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ ПО ЭМОЦИОНАЛЬНОЙ КРИВОЙ СЮЖЕТА	35
--	----

<i>Виноградова Ю.С., Ильченко И.В., Ширяева П.С.</i> АВТОМАТИЧЕСКОЕ АННОТИРОВАНИЕ ОШИБОК В УСТНОЙ РЕЧИ НА АНГЛИЙСКОМ ЯЗЫКЕ С ПРИМЕНЕНИЕМ LLM	37
<i>Малашенко Б.Т.</i> ДИСТИЛЛЯЦИЯ НА УРОВНЕ ТЕКСТОВ: КАК СОЗДАТЬ МАЛЕНЬКИЙ, НО ЭФФЕКТИВНЫЙ ЭНКОДЕР ПРЕДЛОЖЕНИЙ	39
<i>Шамигов Ф.Ф.</i> ОПТИМИЗАЦИЯ FINE-TUNING’A: СООТНОШЕНИЕ ПАРАМЕТРОВ МОДЕЛИ И ТОКЕНОВ (НА ПРИМЕРЕ ГЕНЕРАЦИИ НОВОСТНЫХ ЗАГОЛОВКОВ И ЛИДОВ)	40
<i>Пименов И.С.</i> АНАЛИЗ ПРИЁМОВ АРГУМЕНТАЦИИ В ТЕКСТАХ НАУЧНОЙ КОММУНИКАЦИИ	42

ПРОМПТ-ИНЖИНИРИНГ, РАЗРАБОТКА ЧАТ-БОТОВ И ПРОЧИЕ ПРИКЛАДНЫЕ NLP-ЗАДАЧИ

<i>Тельнина Л.Д.</i> ПРОТОТИП ЧАТ-БОТА ДЛЯ ВЛАДИМИРО-СУЗДАЛЬСКОГО МУЗЕЯ-ЗАПОВЕДНИКА	45
<i>Золин А.П.</i> РАЗРАБОТКА ЧАТ-БОТОВ ДЛЯ ПЕРСОНАЛИЗИРОВАННОГО ИЗУЧЕНИЯ ИНОСТРАННЫХ ЯЗЫКОВ: БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ	46
<i>Сергеева М.О., Сытикова Е.А.</i> РАЗРАБОТКА ЧАТ-БОТА ДЛЯ ЭРМИТАЖА: АВТОМАТИЗАЦИЯ ВЗАИМОДЕЙСТВИЯ С ПОСЕТИТЕЛЯМИ МУЗЕЯ	48
<i>Ващенко Д.В., Куценко А.В.</i> НЕЙРОСЕТЕВОЕ OCR-РАСПОЗНАВАНИЕ ИСТОРИЧЕСКИХ ТЕКСТОВ (НА МАТЕРИАЛЕ ГАЗЕТЫ «ТОМСКИЙ ВЕСТНИК»)	50
<i>Андреевская Е.А., Фомин В.Е., Стрыгин Д.А.</i> ИССЛЕДОВАНИЕ И ВЫБОР МОДЕЛИ ФИЛЬТРАЦИИ ТОКСИЧНЫХ СООБЩЕНИЙ ДЛЯ ЦИФРОВОГО АССИСТЕНТА ОБРАЗОВАТЕЛЬНОЙ ПЛАТФОРМЫ	51
<i>Бирюков В.М.</i> ПРОМПТ-ИНЖИНИРИНГ И СОВЕРШЕНСТВОВАНИЕ ВЗАИМОДЕЙСТВИЯ С ИНСТРУМЕНТАМИ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА	53

Научное издание

**ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА
ДЛЯ ЛИНГВИСТИКИ, IT И БИЗНЕСА**

**Сборник материалов
Международной научно-практической конференции
(28 февраля – 1 марта 2025 г.)**

Оригинал-макет А.И. Лелоюр
Дизайн обложки Л.Д. Кривцовой

Подписано в печать 24.07.2025 г. Формат 60×84¹/₁₆.
Печ. л. 3,6. Усл. печ. л. 3,3. Гарнитура Times.
Тираж 500 экз. Заказ № 6400.

Отпечатано на полиграфическом оборудовании
Издательства Томского государственного университета
634050, г. Томск, пр. Ленина, 36.
Телефоны: 8(382-2)–52-98-49; 8(382-2)–52-96-75
Сайт: <http://publish.tsu.ru>
E-mail: rio.tsu@mail.ru

ISBN 978-5-907890-87-9



