УДК 81.322.2

Русскоязычное автоматическое реферирование: можно ли решить проблему ограниченности данных архитектурой? Ахметгареева А., Абрамов А., Кулешов И., Лещук В., Феногенова А. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 5–26.

В данной работе исследуется проблема автоматического реферирования, акцентируя внимание на её значимость, вызовы и методы, особенно в контексте русского языка. Мы выделяем ограничения текущих метрик оценки и наборов данных, которые представляют различные сценарии реферирования. В работе изучены различные подходы, включая форматы контролируемого обучения, сравнение моделей, предназначенных для русского языка, и обладающих кросс-языковыми возможностями, а также влияние настройки обучения с подкреплением на конечные результаты. Вклад работы включает изучение задачи реферирования для русского языка, публикацию набора данных на основе инструкций и лучшей открытой модели, а также перспективы для дальнейших достижений в данной области.

Библ. – 43 назв.

УДК 81.322.2

Улучшение совместных вложений текстов и кода для задачи поиска с эффективным по параметрам дообучением. Галлямов К., Хаертдинова Л., Денисова К. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 27–45.

Последние достижения в области обработки естественного языка (NLP) демонстрируют значительный прогресс в задаче поиска по исходному коду. По мере увеличения размеров моделей на базе трансформеров, используемых в этой задаче, возрастают вычислительные затраты и время, необходимые для полного их дообучения. Это представляет серьёзную проблему для адаптации и использования этих моделей в условиях ограниченных вычислительных ресурсов. В связи с

этими проблемами мы предлагаем метод дообучения, который использует техники эффективного по параметрам дообучения (PEFT). Кроме того, мы применяем контрастивные функции ошибки для улучшения качества бимодальных представлений, обучаемых моделями на основе трансформеров. Для методов PEFT мы предоставляем широкие сравнительные оценки, отсутствие которых было отмечено как важная проблема в литературе. На основе экспериментов с моделью CodeT5+, проведённых на двух наборах данных, мы демонстрируем, что предложенный фреймворк настройки способен улучшить эффективность поиска по коду и тексту, настраивая не более 0.4% параметров.

Библ. – 25 назв.

УДК 004.852

UnGAN: методы машинного разучивания через атаку на наличие в обучающей выборке. Жаворонкин А., Паутов М., Калмыков Н., Севрюгов Е., Ковалев Д., Рогов О. Ю., Оселедец И. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 46–60.

В условиях растущих требований к конфиденциальности данны46х и праву на забвение, способность эффективно исключать определенные данные из моделей машинного обучения без повторного обучения с нуля становится решающей. Машинное разучивание направлено на эффективное устранение влияния некоторых данных на модель. Мы предлагаем **UnGAN**, новый подход к машинному разучиванию, использующий генеративно-состязательные сети (GAN) для удовлетворения растущей потребности в эффективном и надежном удалении данных из обученных моделей. UnGAN предлагает уникальную стратегию разучивания через атаку на наличие в обучающей выборке, где дискриминаторная сеть обучается определять, был ли данный ввод частью набора данных для обучения модели. Дискриминатор представляет собой трехслойную полностью соединённую сеть с функциями активации ReLU, принимающую входы от вывода модели, подвергающейся разучиванию, и метку класса. Эта архитектура позволяет дискриминатору с высокой точностью определять статус членства данных, что позволяет управлять процессом разучивания.

Библ. – 35 назв.

УДК 81.322.2

Эффективность методов извлечения персон. Зайцев К. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 61–81.

В работе представлен анализ методов извлечения информации о участниках диалога и оценка их производительности на русском языке. Для обучения моделей для данной задачи набор данных Multi-Session Chat был переведен на русский язык с использованием нескольких моделей перевода, что привело к улучшению качества данных. Представлена метрика, основанная на концепции F-меры, для оценки эффективности моделей извлечения. Метрика использует обученный классификатор для определения участника диалога, которому принадлежит персона. Эксперименты проводились на моделях MBart, FRED-T5, Starling-7B, основанной на Mistral, и моделях Encoder2Encoder. Результаты показали, что все модели продемонстрировали недостаточный уровень полноты в задаче извлечения персон. Включение функции NCE Loss улучшило точность модели за счет уменьшения полноты. Кроме того, увеличение размера модели привело к улучшению извлечения персон.

Библ. – 31 назв.

УДК 004.855.5

Оптимизация конвейера разработки признаков в AutoML с использованием крупных языковых моделей. Иов И. Л., Никитин Н. О. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 82–112.

Одним из важных путей достижения более эффективного автоматизированного машинного обучения является применение мета-оптимизации на всех этапах проектирования конвейера. В данной работе мы стремимся использовать крупные языковые модели для этапов разработки признаков как в роли оптимизаторов, так и экспертов в области знаний. Мы закодировали конвейер разработки признаков на естественном языке в виде последовательности атомарных операций. “Черный ящик” оптимизации реализован путем запроса конвейера разработки признаков у языковой модели с использованием подсказки, состоящей из предопределенных инструкций, описания набора данных и ранее оцененных конвейеров. Для увеличения временной эффективности и стабильности оптимизации был реализован алгоритм на основе популяций, генерирующий набор конвейеров с каждым ответом

языковой модели вместо одного. Совместно было предложено многократное улучшение, чтобы предоставить языковой модели дополнительную доменную информацию. Чтобы проанализировать применимость предложенного подхода, мы проводим серию экспериментов на открытых наборах данных. В качестве базового подхода для задачи оптимизации был выбран метод случайного поиска. Прямые результаты, полученные с использованием модели gpt-3.5-turbo, близки к базовому подходу с той же временной стоимостью. Генерация конвейеров на основе популяций превосходит базовый подход и другие методы. Это подтверждает, что предложенный подход может повысить общую производительность моделей машинного обучения при таких же временных затратах на оптимизацию и меньшем количестве токенов для получения результата.

Библ. – 62 назв.

УДК 004.932.4

Ti-Patch: плиточный физический адверсиальный патч для метрик качества видео без эталона. Леоненкова В., Шумитская Е., Анциферова А., Ватолин Д. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 113–131.

Объективные метрики качества изображений и видео без эталона играют ключевую роль во многих задачах компьютерного зрения. Однако современные метрики без эталона стали основываться на обучении и уязвимы перед состязательными атаками. Уязвимость метрик качества накладывает ограничения на использование таких метрик в системах контроля качества и при сравнении объективных алгоритмов. Кроме того, использование уязвимых метрик в качестве функции потерь при обучении моделей глубокого обучения может привести к ухудшению визуального качества. В связи с этим тестирование метрик качества на уязвимость является актуальной задачей. В настоящей работе предлагается новый метод тестирования уязвимости метрик качества в физическом пространстве. Насколько нам известно, метрики качества ранее не тестировались на уязвимость к такой атаке; они тестировались только в пиксельном пространстве. Мы применили физическую состязательную атаку Ti-Patch — плиточный патч — к метрикам качества и провели эксперименты как в пиксельном, так и в физическом пространстве. Также мы провели эксперименты по реализации физических состязательных “обоев”. Предложенный метод может быть

использован как дополнительная метрика качества при оценке уязвимости, дополняя традиционные субъективные сравнения и тесты уязвимости в пиксельном пространстве. Мы разместили наш код и адверсиальные видео на GitHub: `https://github.com/leonenkova/Ti-Patch`.

Библ. – 39 назв.

УДК 004.852

Оценка максимального уровня шума в задачах черного ящика. Лобанов А., Гасников А. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 132–147.

В задачах оптимизации черного ящика точная оценка максимального уровня шума имеет ключевое значение для обеспечения надежной работы. В настоящей работе предлагается новый подход к улучшению оценки максимального уровня шума, сосредоточенный на сценариях, где доступны только значения функции, возможно, с ограниченным состязательным шумом. Используя безградиентные алгоритмы оптимизации, мы вводим новое ограничение на шум, основанное на предположении о липшицевости, что позволяет улучшить оценку уровня шума (или улучшить уровень ошибки) для негладких и выпуклых функций. Теоретический анализ и численные эксперименты демонстрируют эффективность нашего подхода, даже для гладких и выпуклых функций. Данное достижение способствует повышению надежности и эффективности алгоритмов оптимизации черного ящика в различных областях, таких как машинное обучение и проектирование инженерных систем, где состязательный шум представляет значительную проблему.

Библ. – 32 назв.

УДК 004.942

Выявление и устранение ковариантных сдвигов в данных для более надежного прогнозирования отказов HDD. Лукьянов К., Дробышевский М., Турдаков Д. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 148–161.

Прогнозирование отказов жестких дисков (HDD) привлекает значительное внимание в исследованиях, однако наличие сдвигов переменных (covariate shifts) в данных остается практической проблемой. В данном исследовании предлагается новый подход к обучению моделей

обнаружения сдвигов переменных без необходимости дополнительного использования реальных данных или моделирования искусственных сдвигов. Кроме того, предлагается комплексная методология, интегрирующая обнаружение сдвигов, уведомления администраторов, устранение сдвигов и прогнозирование отказов HDD. Экспериментальные результаты демонстрируют жизнеспособность предлагаемого метода для реального применения.

Библ. – 39 назв.

УДК 81.322.2

Улучшение RAG с помощью дообучения LoRA для генерации текста персонажа. Павлюкевич В., Жердева А., Махныткина О., Дырмовский Д. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 162–177.

В статье рассматривается задача поддержания согласованности в системах порождения текста с использованием поиска (Retrieval-Augmented Generation, RAG) для порождения текста персонажей в случаях, когда базы данных подвержены частым обновлениям, и стандартное дообучение больших языковых моделей (LLM) оказывается недостаточной. Мы предлагаем подход, который улучшает существующую систему RAG, применяемую для поиска информации, основанной на персонажах в диалоговых агентах, посредством дообучения с использованием Low-Rank Adaptation на синтетических данных. Нами было установлено, что данный метод улучшает логику и точность системы на 5% по оценкам SSA и обеспечивает создание более связного и контекстуально релевантного контента.

Библ. – 23 назв.

УДК 004.852

Открытая библиотека для мультимодальной кластеризации методами AutoML на Apache Spark. Муравьёв С., Казаковцев В., Усов И., Шпинева П., Муравьёва О., Шалыто А. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 178–193.

Мы представляем библиотеку, которая позволяет выбирать и настраивать алгоритмы кластеризации для мультимодальных данных, то есть данных, где каждый объект представлен не только вектором,

но также текстом и/или изображением, и каждая модальность значима. Наша библиотека автоматически находит баланс между исследованием и эксплуатацией входных данных среди набора реализованных алгоритмов кластеризации в соответствии с выбранным внутренним индексом валидации кластеризации. В библиотеке также реализована рекомендательная система для выбора индекса валидации, которая может предсказать наиболее подходящую меру для входных данных. Мы использовали Apache Spark для реализации алгоритмов кластеризации, что позволяет использовать библиотеку на распределённых вычислительных системах для кластеризации больших мультимодальных данных.

Библ. – 12 назв.

УДК 81.322.2

Обзор задачи автоматического порождения ответов на юридические вопросы. Сабиева А., Жаманхан А., Жетесов Н., Кубаева А., Ахметов И., Пак А., Ахметова Д., Жаксылыкова А., Еленов А. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 194–213.

Недавние достижения в области многодокументной суммаризации в юридической сфере демонстрируют значительный прогресс в извлечении и сокращении информации из юридических текстов. Современные методы используют комбинацию обработки естественного языка, машинного обучения и методов добычи данных для выявления и выделения ключевых элементов и тем из множества юридических документов. Этот процесс создает структурированные, краткие и релевантные обзоры на основе конкретных юридических запросов или тем, часто называемых многодокументными рефератами. Такие рефераты помогают более эффективно фиксировать суть сложных и обширных юридических материалов без потери необходимой детализации. Основное внимание в последних исследованиях уделяется повышению точности извлечения информации, улучшению связности созданных обзоров и обеспечению актуальности содержания конкретной юридической проблематыке. Хотя проблемы ещё остаются, особенно в нюансах юридического языка и разнообразии типов документов, направление развития области движется в сторону более сложных и удобных для пользователя систем, что обещает существенную трансформацию юридических исследований и доступности информации.

Библ. – 23 назв.

УДК 81.322.2

ММА: борьба за ускорение многоязыковых моделей. Сухановский Н., Рындин М. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 214–232.

В работе мы рассматриваем стандартный способ проектирования моделей для обработки естественного языка: дообучение многоязыковой языковой модели, в котором данные для целевой задачи на одном языке используются для последующего решения этой задачи на другом целевом языке. Цель работы – определить, как популярные методы ускорения моделей машинного обучения влияют на многоязыковые возможности моделей на основе трансформеров, а также исследуем использование этих методов в различных комбинациях. В результате мы получаем модель NERC, которая может эффективно работать на CPU и сохраняет многоязыковые свойства для нескольких тестовых языков после настройки и ускорения только с использованием данных на английском языке.

Библ. – 24 назв.

УДК 81.322.2

Как обеспечить надежный код: использование статического анализатора для выявления и устранения дефектов в автоматически порожденном коде. Шайхелисламов Д., Дробышевский М., Белеванцев А. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 233–251.

Развитие больших языковых моделей (LLM) значительно расширило возможности порождения кода. Недавний опрос на StackOverflow показал, что 70% разработчиков используют или планируют использовать инструменты ИИ в разработке кода. Однако большинство существующих методов сосредоточены на задачах дообучения с учителем, заимствованных из порождения текстов, что часто упускает такие важные особенности кода, как возможность компиляции и синтаксическая и функциональная корректность. Для решения этой проблемы мы предлагаем новый подход, сочетающий предобученные LLMs с инструментами анализа программного обеспечения, которые широко используются для обнаружения уязвимостей и проверки кода. Наш метод использует подробную обратную связь от компилятора и инструментов анализа кода, интегрируя эти специализированные знания в

процесс порождения подсказок. Мы представляем CodePatchLLM, расширение больших языковых моделей, использующее Svace для улучшения порождения кода. Это универсальный фреймворк, поддерживающий несколько языков программирования. Обширное экспериментальное исследование на наборе данных LeetCode показывает, что наш подход превосходит базовую модель CodeLlama, значительно улучшая показатели успешности компиляции и функциональной корректности для Java, Python и Kotlin. Код CodePatchLLM доступен по адресу `https://github.com/dsshay/CodePatchLLM`.

Библ. – 55 назв.

УДК 81.322.2

Применение синтаксических парсеров для турецкого языка в задаче разметки кыргызских синтаксических корпусов. Алексеев А., Тиллабаева А., Кабаева Г. Дж., Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 252–275.

Кыргызский (киргизский) язык, как один из малоресурсных, требует значительных усилий для создания качественных синтаксических корпусов. В данной работе предложен вариант подхода, упрощающего процесс разработки синтаксического корпуса для кыргызского языка. В настоящей работе представлен инструмент для переноса синтаксической разметки с турецкого языка на кыргызский, основанный на методе машинного перевода трибанков. Эффективность предложенного инструмента была оценена с использованием трибанка TueCL. Результаты исследования показывают, что данный подход обеспечивает более высокую точность синтаксической разметки по сравнению с моноязычной моделью, обученной на кыргызском трибанке KTMU. Кроме того, в работе предлагается метод оценки сложности ручного аннотирования полученных синтаксических деревьев.

Библ. – 45 назв.

УДК 81.322.2

Применения больших языковых моделей для задач порождения и обработки программного кода. Ломшаков В. М., Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 276–350.

В последние годы большие языковые модели (Large Language Models, LLM) существенно изменили подходы к автоматизации программирования, предоставляя мощные инструменты для порождения, исправления и оптимизации кода. В настоящем обзоре мы рассматриваем методы адаптации LLM к задачам программирования, включая обучение с подкреплением на основе человеческих предпочтений (RLHF), дообучение следованию инструкциям (instruction tuning), адаптивные подходы (PEFT) и эффективные стратегии промптинга (prompting). Мы систематизируем современные методы дообучения и применения моделей, обсуждаем их преимущества и ограничения, рассматриваем актуальные датасеты для задача порождения и исправления кода и метрики их оценивания, а также описываем передовые модели с открытыми весами для работы с кодом.

Библ. – 145 назв.

УДК 004.852

Непараметрические методы решения задачи согласования данных. Гаранин В. А., Семенов К. К. — В кн.: Исследования по прикладной математике и информатике. IV. (Зап. научн. семин. ПОМИ, т. 540), СПб., 2024, с. 351–405.

В работе рассмотрено текущее состояние задачи согласования индустриальных данных, а также основные подходы к её решению как в классических, так и в неклассических постановках. Представлен всесторонний обзор основных направлений в данной области и предложены новые подходы, среди которых — методы непараметрического согласования данных, а также подход, который возвращается к аналитическим решениям, но на современном уровне сложности, требований и ограничений, накладываемых на конечные результаты задачи согласования. Метод предлагает замкнутые формулы, позволяющие проводить независимую и индивидуальную оптимизацию процесса согласования индустриальных данных для конкретных условий, что ранее в научной литераткуре не достигалось. Данная работатакже нацелена на восполнение существующего пробела в научной литературе на русском языке по тематике согласования индустриальных данных.

Библ. – 82 назв.