

Тардиф А. Отчет LXT «Путь к зрелости искусственного интеллекта в 2024 г.»: раскрываем будущее инноваций в области искусственного интеллекта и корпоративной трансформации // Unite. AI. 2024. 23 апр. URL: <https://www.unite.ai/ru/путь-к-зрелости-ИИ-в-2023-году/> (дата обращения: 10.04.2024).

Санамян Л. ТОП-10 нейросетей для поиска информации в интернете // Otzyvmarketing. 2023. 18 сент. URL: <https://otzyvmarketing.ru/articles/neiroseti-dlya-poiska-informacii-v-internete/> (дата обращения: 10.04.2024).

Что такое искусственный интеллект // Garpix. 2023. 14 марта. URL: <https://garpix.com/blog/chto-takoe-iskusstvennyj-intellekt> (дата обращения: 13.04.2024).

УДК 004.932.72:1:002.1-028.21

И. А. Волегов¹

Российский государственный профессионально-педагогический университет

РАСПОЗНАВАНИЕ ТЕКСТОВ С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. В работе представлены способы и проблемы распознавания текстов из старинных и антикварных книг, архивных материалов и редких документов. Освещаются трудности, с которыми сталкиваются традиционные методы оптического распознавания символов (OCR), при работе с различными стилями письма, устаревшими словами и разнообразными форматами документов. Современные методы OCR на основе нейронных сетей, таких как сверточные нейронные сети и рекуррентные нейронные сети, которые способны эффективно обрабатывать текст с различными искажениями и стилями.

Ключевые слова: OCR, оцифровка текста, распознавание текста, нейронные сети, сверточные нейронные сети, рекуррентные нейронные сети, LSTM.

Проблема сохранения культурного и исторического наследия остро встает перед необходимостью распознавания и оцифров-

¹ Научный руководитель: С. В. Ляхов, кандидат технических наук, доцент, заведующий кафедрой РГППУ.

ки текстов из старинных и антикварных книг, архивных материалов и редких документов. Важность этого процесса проявляется особенно в случае редких и ценных книг, которые из-за своего состояния или возраста могут быть трудночитаемыми и ветхими по состоянию. Такие материалы требуют специализированного подхода к распознаванию текста, учитывая их уникальные особенности и потенциальные трудности в считывании [Шипилин].

Оцифровка и распознавание текста этих материалов представляет собой существенное решение проблемы их сохранения. В библиотечных и архивных фондах особенно важна оцифровка, так как это позволяет не только обеспечить доступ к содержимому документов в цифровом виде, но и сохранить их оригинальные копии от повреждений и утраты. Для библиотек акцент делается на доступности информации для пользователей, а оцифровка играет ключевую роль в этом процессе, расширяя доступ к содержимому старинных и ценных изданий [Нуждва, с. 60].

Если принять во внимание тот факт, что при Московском государственном институте международных отношений открыт Музей Редкой книги, некоторые экземпляры которого могут быть повреждены или иметь нечитаемый текст, то значимость оцифровки и распознавания возрастает.

Создание полнотекстовых баз данных (БД) на основе распознанного текста редких книг и архивных документов позволяет формировать цифровые коллекции, доступные для исследований и использования. Такие БД обеспечивают долгосрочное сохранение и удобный доступ к уникальному материалу для будущих поколений, сохраняя его целостность и уникальные особенности.

Цель работы — повышение эффективности распознавания старинных и уникальных текстовых рукописных материалов.

Старинные книги и документы могут быть повреждены, иметь износ, загрязнения, засветы или другие дефекты, которые усложняют распознавание текста. Неровные края страниц, потертости поверхности страниц или размытие, также, могут затруднить процесс оцифровки и распознавания.

Исходные материалы могут содержать различные стили рукописного или печатного текста, которые не всегда легко распознать для традиционных алгоритмов OCR. Например, старинные рукописи или шрифты редко встречающихся стилей могут быть сложными для автоматического распознавания и содержать устаревшие слова, архаичные фразы или специфические языковые конструкции, которые могут быть непонятны современным системам распознавания текста.

К тому же, часто встречаются нестандартные или неоднородные форматы, например, изменяющиеся стили и размеры шрифтов, различные межстрочные интервалы или расположение текста на странице. Это может усложнить точное выделение и распознавание текстовой информации.

При распознавании старинных документов и книг важно учитывать и сохранять исторический и культурный контекст, который может быть важным для понимания текста. ИИ может столкнуться с трудностями в интерпретации и передаче такого контекста.

Для исторических и редких материалов требуется высокая точность распознавания текста. Ошибки при распознавании могут привести к потере важной информации или искажению смысла [Как сканировать книги].

Распознавание текста из изображения или же оптическое распознавание символов (англ. optical character recognition, OCR) — это технология для автоматизации извлечения данных из печатного, письменного текста, отсканированного документа, файла изображения с целью последующего преобразования текста в машиночитаемую форму. Данная форма будет использована для работы с данными, например, редактирование или поиск информации [Бобров, с. 124].

Каждая система OCR состоит из следующих шагов алгоритма:

Предобработка.

Сегментация.

Выделение признаков.

Распознавание символов или классификация.

Постобработка и исправления ошибок распознавания.

Эти алгоритмические шаги выполняются последовательно, и каждый результат шага подается на вход следующего шага.

Главное отличие всех методов распознавания текста заключается на этапе классификации изображения. Классификация позволяет распознать символ из изображения и перевести его в машиночитаемый формат. Существуют разные методы распознавания, самыми популярными являются [Вальке]:

Шаблонный.

Признаковый.

Нейросетевой.

Шаблонный метод основан на идее сравнения изображенных символов с заранее известными шаблонами символов (так называемыми «образцами» или «шрифтами»). Каждый символ на изображении сравнивается с библиотекой или базой данных шаблонов, и выбирается наиболее похожий шаблон. Наиболее подходящим шаблоном считается тот, у которого будет наименьшее количество точек, отличных от исследуемого изображения.

Признаковый подход использует статистические методы для анализа текста. Например, может применяться модель языка, которая оценивает вероятность появления определенной последовательности символов или слов на основе статистики из обучающих данных.

Главные достоинства статистических методов — это простота их реализации, хорошая устойчивость к изменениям формы символов, низкое количество ошибок при распознавании, высокое быстродействие.

Самые главные недостатки данного метода — неустойчивость к различным дефектам изображения, например, шум, а также на этапе извлечения признаков из символа происходит потеря основной информации, извлечение ведется независимо, из-за чего расположение элементов символа утрачивается.

Нейронные сети являются современным и эффективным подходом к распознаванию текста с использованием OCR. Обычно для OCR применяются сверточные нейронные сети (CNN) или рекуррентные нейронные сети (RNN), такие как LSTM (Long Short-Term Memory). Нейронные сети обучаются на больших наборах данных изображений текста, где они автоматически изучают признаки символов и их связи, что позволяет им распознавать текст даже при наличии искажений, различных стилей письма или плохого качества изображений.

Основные недостатки — текст должен быть в вертикальном положении, сложность подбора обучающей выборки. Основные достоинства — это высокая скорость и обобщённость. Именно поэтому данный метод сейчас используется в разных современных системах распознавания текста.

Таким образом, нейросетевые методы, особенно с использованием сверточных и рекуррентных нейронных сетей, будут эффективны по следующим причинам:

- способность обучаться сложным признакам и структурам в изображениях текста;

- высокая устойчивость к различным формам искажений и шума на изображениях

- эффективное распознавание текста даже при наличии различных стилей письма и форм символов

Список источников и литературы:

Шипилин П. И., Назарчук Т. Б. Актуальные проблемы оцифровки артефактных, редких и ценных документов на примере научной библиотеки Таврической Академии КФУ и научного архива ГИАМЗ «Херсонес Таврический» // Научный вестник Крыма. 2017.

№ 1 (6). С. 110–115. URL: <https://cyberleninka.ru/article/n/aktualnye-problemy-otsifrovki-artefaktnyh-redkih-i-tsennyh-dokumentov-na-primere-nauchnoy-biblioteki-tavricheskoy-akademii-kfu-i/viewer> (дата обращения: 25.03.2024).

Нуждова Д. А. Нейросети в библиотечном деле: опыт проекта «Новые библиотекари» // Корпоративные библиотечные системы: технологии и инновации (КорФор-2023): материалы Международной научно-практической конференции и выставки, 19–23 июня 2023 г. СПб.: ПОЛИТЕХ-ПРЕСС, 2023. С. 59–65. <https://doi.org/10.18720/SPBPU/2/k23-6>.

Бобров К. А., Шульман В. Д., Власов К. П. Анализ технологий распознавания текста из изображения // Международный журнал гуманитарных и естественных наук. 2022. № 3 (66), ч. 2. С. 124–128. <https://doi.org/10.24412/2500-1000-2022-3-2-124-128>.

Притула П. Как сканировать книги и каталоги библиотеки: технологии и практика // СиНьюс: издание в сфере корпоративных информационных технологий в России и странах СНГ. 2016. 29 янв. URL: https://scan.cnews.ru/articles/2016-01-29_skanirovanie_knig_i_kataloga_biblioteki_tehnologii_i_praktika (дата обращения: 25.03.2024).

Вальке А. А., Лобов Д. Г. Алгоритмы распознавания символов // Динамика систем, механизмов и машин. 2018. Т. 6, № 4. С. 164–168. URL: <https://cyberleninka.ru/article/n/algoritmy-raspoznavaniya-simvolov/viewer> (дата обращения: 20.03.2024).

УДК 005.935:004.896

С. А. Гирич¹

Российский государственный профессионально-педагогический университет

КОНТРОЛЬ ВЫПОЛНЕНИЯ ПРОИЗВОДСТВЕННЫХ ЗАДАЧ

Аннотация. В работе анализируется потенциал цифровых помощников и систем на базе искусственного интеллекта для контроля выполнения производственных задач, разрабатываются рекомендации по оптимальному использованию цифровых помощников. Подтверждается значительный потенциал таких помощников для автоматизации, увеличения производительности и обеспечения безопасности труда.

¹ Научный руководитель: С. В. Ляхов, кандидат технических наук, доцент, заведующий кафедрой РГППУ.