УДК 004.82

РАЗРАБОТКА АЛГОРИТМА НОРМАЛИЗАЦИИ РАСПОЗНАННОГО ТЕКСТА OCR-МОДЕЛЯМИ ДЛЯ СТРОГО ФОРМАЛИЗОВАННЫХ ДАННЫХ

Вершинин В.К. (ИТМО)

Научный руководитель – кандидат технических наук, ст. преподаватель Ходненко И.В. (ИТМО)

Введение. В современных организациях активно внедряются электронные архивы, где ключевую роль играет технология OCR (Optical Character Recognition). Однако даже высокоточные OCR-модули зачастую совершают характерные ошибки при распознавании символов: путаница букв и цифр, неправильная сегментация слов, появление «шума» при работе с повреждёнными документами [1]. Для строго формализованных данных (например, реквизиты, технические названия, коды товаров) такие ошибки могут приводить к сбоям в аналитических системах, так как даже единичная подмена символа нарушает автоматическую обработку. Одним из эффективных способов решения проблемы является автоматическая нормализация распознанного текста. Она подразумевает исправление символов, удаление «лишних» фрагментов и восстановление структуры без утраты смысла [2]. Подход с использованием современных языковых моделей (LLM) даёт возможность адаптироваться к различным типам ошибок и поддерживать заданный формат данных. Целью данной работы является разработка «лёгкого» алгоритма нормализации OCR-текстов, который повышает точность распознавания и при этом не требует чрезмерных вычислительных ресурсов.

Основная часть. Предложенный алгоритм сочетает несколько ключевых шагов. Сначала выполняется фильтрация «шума»: очистка от неверных пробелов, повторяющихся символов и иных артефактов. Затем применяется лингвистический анализ для поиска типичных OCR-ошибок, включая неверные подстановки (например, «0» и «O») и пропуски знаков. На следующем этапе используется компактная языковая модель, способная выполнять исправление ошибок по контексту. В работе протестированы версии моделей LLaMA и Mistral, адаптированные с помощью небольшого набора примеров (few-shot prompting) [3]. Это позволяет модели быстро ориентироваться в распространённых искажениях текста и повышать точность без глубокой дорогостоящей дообученности.

Для оценки эффективности применяются классические метрики: доля исправленных символов и слов (Character Error Rate, Word-Level Accuracy), а также показатель семантической близости. Полученные результаты демонстрируют заметное повышение корректности распознанного текста (сокращение ошибок на 5-7% по сравнению с исходными выводами OCR), особенно если исходные данные содержат символы, где буквы и цифры визуально похожи. Предварительные эксперименты также показывают, что дополнительное использование правил валидации (например, регулярных выражений) улучшает качество итоговой нормализации при наличии строго определённого формата.

Таким образом, оптимальное решение задачи автоматической нормализации OCR-текстов может быть получено путём комбинированного подхода, когда расширенные языковые модели исправляют смысловые искажения, а простые правила и словари убирают типовые подстановки и контролируют формат данных [4].

Выводы. Разработан алгоритм нормализации распознанного текста, позволяющий исправлять типичные OCR-ошибки и восстанавливать структуру документов для формализованных данных.

Использование компактных языковых моделей в сочетании с эвристическими фильтрами шума даёт ощутимый прирост точности при умеренном уровне вычислительных затрат.

Эксперименты подтверждают целесообразность гибридного подхода, особенно при работе с текстами, где подмена символов критична для корректной дальнейшей

автоматизации.

В дальнейших исследованиях планируется расширение набора обучающих примеров и тестирование на отраслевых дата-сетах с разным уровнем шума.

Список использованных источников:

1. Doshi F., Gandhi J., Gosalia D., Bagul S. Normalizing Text Using Language Modelling Based on Phonetics and String Similarity // arXiv.org, 2020.
2. Bitton Y., Ro J., Ash R., Pfeffer S. Adversarial Text Normalization // arXiv.org, 2022.
3. Bollmann M., Bingel J., Søgaard A. A Large-Scale Comparison of Historical Text Normalization Systems // arXiv.org, 2019.
4. Bocur C., van der Goot R. Lexical Normalization Based on Multilingual Transformers // arXiv.org, 2021.

Автор _____ Вершинин В.К.

Научный руководитель _____ Ходненко И.В.