Распознавание и поиск по рукописным текстам и их слабой разметке

Зыков Валерий Павлович, Морозов Иван Дмитриевич

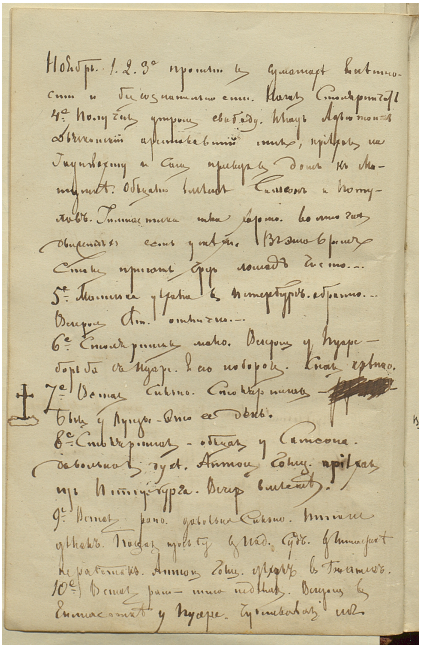
Студент, студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: zykovvp@my.msu.ru, morozov-ivan-2003@yandex.ru

Научный руководитель *— Местецкий Леонид Моисеевич*

В работе рассматривается задача распознавания рукописных текстов архива А. В. Сухово-Кобылина, важного исторического источника XIX века, и последующий поиск по полученной расшифровке с ошибками. Основная сложность задачи заключается в плотной компоновке строк, множестве зачеркиваний, пропусков символов и многоязычных вставках. Основной целью исследования является разработка метода, который сможет эффективно распознавать текст, несмотря на указанные особенности, а также методов поиска слов и именованных сущностей в данном тексте.



Пример страницы дневника А. В. Сухово-Кобылина

Для решения задачи использовался строчный вариант сверточной нейронной сети *Vertical Attention Network* (VAN), описанной в [2]. Применение строчной модели оправдано тем, что модель End-to-End, работающая на уровне всей страницы, оказалась неэффективной для данных с нелинейными строками и плотной компоновкой текста. Процесс распознавания включает несколько этапов: предварительная сегментация строк на основе алгоритмов, описанных в [1], распознавание отдельных строк с помощью модели VAN [2], а затем пост-обработка результатов с применением языковой модели ChatGPT для исправления явных ошибок. Далее производился поиск слов и именованных сущностей по полученной расшифровке: использовались как классические подходы на основе расстояния Левенштейна, так и поиск с помощью больших языковых моделей, таких как ChatGPT, DeepSeek.

Основные этапы работы:

- Сегментация строк на страницах с помощью программы, разработанной Л. М. Местецким.
- Обучение модели VAN на предварительно размеченных данных (IAM dataset [3]) с последующим дообучением на архиве Сухово-Кобылина.
- Исправление ошибок распознавания с помощью ChatGPT, использующей подход *few-shot learning*.
- Поиск слов по запросу и поиск именованных сущностей, используя полученную слабую расшифровку с ошибками.

Эксперименты и результаты: Модель обучалась на размеченных данных архива, состоящих из 1505 строк для тренировки, 119 строк для валидации и 74 строк для тестирования. В ходе экспериментов было проведено несколько подходов к улучшению качества и организации поиска:

- Обучение модели на исходных строках в формате RGB дало метрики CER = 19.03% и WER = 52.44%.
- Бинаризация строк с удалением свисающих элементов букв ухудшила результаты (CER = 25.20%, WER = 66.77%).

- Использование ChatGPT для пост-обработки позволило снизить ошибку до CER = 17.98% и WER = 41.32%, что является значительным улучшением.

- Поиск по запросу дал пропуск цели в 5.23% при использовании top-50 результатов.

- Поиск именованных сущностей с помощью DeepSeek позволил выделить основные даты, локации и персоны, соотносящиеся с экспертной разметкой.

Заключение: Полученные результаты показывают, что предложенный метод, основанный на сегментации строк и распознавании с применением модели VAN, является эффективным для работы с архивами Сухово-Кобылина. Однако наибольшее улучшение удалось достичь с использованием языковой модели ChatGPT, что позволяет предположить, что комбинированный подход компьютерного зрения и языковой обработки является перспективным направлением для дальнейших исследований. В будущем планируется привлечения аппарата штрихового разложения для выделения новых признаков и совершенствования алгоритмов поиска, а также улучшение сегментации строк и дальнейшие эксперименты с языковыми моделями для исправления ошибок распознавания.

Работа поддержана грантом РНФ №22-68-00066 «Культурное наследие России: интеллектуальный анализ и тематическое моделирование корпуса рукописных текстов».

Литература

1. Местецкий Л. М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. М.: Физматлит, 2009.
2. Coquenet D., Chatelain C., Paquet T. End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, V. 45, №1, P. 508-524, doi: 10.1109/TPAMI.2022.3144899
3. Marti U.-V., Bunke H. The IAM-database: an English sentence database for offline handwriting recognition, IJDAR, 2002.