**УДК 004.622; 004.912; 004.932.75'1**

# РАЗРАБОТКА ПРОГРАММЫ ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА ОБРАБОТКИ ИНФОРМАЦИИ БИБЛИОГРАФИЧЕСКИХ КАРТОЧЕК

**Коробковский В.А.** (Университет ИТМО)

Научный руководитель – к.т.н., доцент Горлушкина Н.Н.

(Университет ИТМО)

**Введение.** В XXI веке прослеживается четкая тенденция на проведение цифровизации. Исключением не стали и библиотеки, которые активно занимаются созданием электронных каталогов из информации, хранящихся в их фондах. Наиболее популярной системой автоматизации, предназначенной для создания и ведения электронной библиотеки, является ИРБИС64+ [1]. Библиографические данные добавляются в систему на основе коммуникативного формата RUSMARC, являющегося национальной адаптацией формата UNIMARC. Однако на текущий момент этот процесс выполняется вручную, что занимает большое количество времени и сил, поэтому конечной целью стала разработка программы для его автоматизации с сопутствующим изучением возникающих проблем и предложением вариантов для их преодоления.

**Основная часть.** Работа ведётся с отсканированными изображениями (карточками) с библиографическим описанием. На вход программе подаётся архив с карточками, а в результате её работы получается файл с систематизированной в формате RUSMARC информацией.

В рамках предобработки были проведены следующие этапы:

1) Удаление чёрных границ, образующихся в результате сканирования.
2) Поворот изображений для выравнивания текста по горизонтали.
3) Применение билатерального фильтра для удаления шумов на изображениях, чтобы текст более четко определялся в процессе распознавания.
4) Перевод изображений из оттенков серого в бинарные [2], чтобы чёрный текст был легко различим на белом фоне.

Каждая карточка имеет определенную структуру, а именно поля автора, заглавия и шифра и условий хранения. В процессе локализации необходимо было разделить изображение на эти поля. Библиотека для оптического распознавания символов pytesseract [3] позволяет просмотреть таблицу с координатами каждого распознанного слова и информацию о нём. С помощью неё строится прямоугольная область, выделяющая поля автора и заглавия на изображениях. Отдельно выделяются части слева и сверху от этой области для получения информации из поля шифра и хранения. Также был создан список стоп-слов для удаления блоков текста с информацией, которая не относится ни к одному из этих полей.

Последним этапом стала работа с полученным текстом. Его необходимо было разбить на соответствующие поля формата RUSMARC. На данном этапе были удалены переносы слов, а также было применено распознавание именованных сущностей для разделения информации из полей автора и заглавия.

Созданная программа была протестирована примерно на 1000 карточек с библиографическим описанием.

**Выводы.** На основании результата работы программного кода можно сказать, что алгоритм локализации дает точность около 80%. Однако в процессе тестирования были отмечены следующие проблемы:

1) Нечеткое распознавание текста. Разумеется, идеального решения не существует, однако всё же можно попытаться использовать другие средств, чтобы сравнить результаты и выбрать лучшее решение.
2) Результат распознавания рукописного текста был либо очень неточным,

либо отсутствовал полностью.

3) Имелось большое количество ошибок при работе с текстом на нескольких языках.

4) Из-за большого разнообразия типов карточек процесс локализации иногда проходил с ошибками, то есть в какую-то область попадала не относящаяся к ней информация.

5) Алгоритм поиска именованных сущностей библиотеки stanza [4] иногда выдавал неправильные результаты, из-за чего разбиение на поля формата RUSMARC происходило с ошибкой.

Таким образом, был сделан вывод о том, что необходимо сократить количество тех операций из алгоритма локализации, которые в теории могут быть выполнены неверно. Возможным решением является использование нейронных сетей, а именно дообучение какой-либо существующей модели для обнаружения текста на собственных примерах. Нейронные сети также можно использовать и для улучшения качества распознавания текста.

**Список использованных источников**:

1) Система ИРБИС64+ [Электронный ресурс]. — URL: https://elnit.org/index.php?option=com_content&view=article&id=255:irbis64&catid=18:kharakteristiki-produktov (дата обращения: 1.10.2023).

2) OpenCV Python Tutorials. Image Thresholding [Электронный ресурс]. — URL: https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html (дата обращения 10.10.2023)

3) Pytesseract [Электронный ресурс]. — URL: https://github.com/madmaze/pytesseract (дата обращения 28.09.2023)

4) Stanza – A Python NLP Package for Many Human Languages [Электронный ресурс]. — URL: https://stanfordnlp.github.io/stanza/ (дата обращения 05.11.2023)