Особенности распознавания текста в архивах дневников Литке

Степочкин Данила Владиславович

Студент

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

E-mail: st.danil.uni@gmail.com

Научный руководитель *— Кропотов Дмитрий Александрович*

В данной работе рассматривается задача обучения моделей распознавания исторических рукописных текстов одного автора, архивов дневников Ф.П. Литке. Дневники адмирала Ф.П. Литке представляют особенный интерес для историков, работающих над периодом России XIX века, но возможность их исследования ограничена необходимостью кропотливой ручной разметки. Эта работа - часть более комплексной задачи по машинной расшифровке документов, включающей в себя также предобработку изображений, сегментацию страниц и строк и постобработку результата.

Архив представляет из себя 4 тома сканированных разворотов текста (более 1000 страниц), часть 2-го и 4-го томов расшифрована (около 300 страниц). Расшифровка и текст содержат характерные особенности, которые необходимо учитывать, такие как дореволюционная орфография, фрагменты на иностранных языках, отдельные неразобранные слова и символы, и другое.

В качестве базовых моделей были использованы архитектуры VAOCR [1] в страничном и строчном варианте, представляющие из себя сверточные сети, обучающиеся с CTC loss.

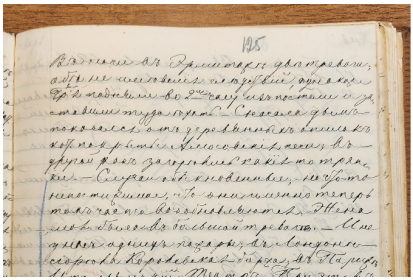
В связи с наличием неразборчивых фрагментов в разметке было необходимо адаптировать стандартную функцию потерь CTC loss. Идея модификации заключается в добавлении специального символа, моделирующего любой другой символ. При этом сохраняется правило CTC, где два одинаковых символа не могут идти подряд, без разделения символом пропуска, в том числе внутри последовательности неопределенных символов. Эта модификация позволяет использовать данные с пропусками фиксированной длины в разметке. Как и с обычным CTC была получена рекуррентная формула. Модификация была проверена в том числе на искусственно зашумленном датасете IAM и было показано что модель может эффективно обучаться (10% CER при 40% замаскированных символов). Также была разработана модификация CTC loss, допускающая пропуски неопределенной длины. Но из-за возрастающей вычислительной

сложности она не была задействована на практике.

В процессе расшифровки также была поставлена задача выделения всех фрагментов на иностранном языке для поиска специальных терминов. Было решено выделять такие фрагменты посимвольно, для чего была сформирована копия датасета с маскированными словами на иностранном языке. Модель была модифицирована для включения в себя нескольких декодеров над единственным энкодером, которые могли обучаться параллельно на разных датасетах. Было показано что если обучать в таком режиме один декодер на обнаружение иностранных символов, а другой на их распознавание, то качество обнаружения повышается (11.27 CER против 13.2 CER на выборке строчек с маскированным иностранным языком).

Обученные строчные модели достигают качества 9% CER для второго тома и 4% CER для четвертого тома, и зачастую даже правильно распознают текст, ошибочно размеченный человеком. Они позволили получить черновую расшифровку для неразобранных частей архива, которая предполагается к использованию в алгоритмах поиска ключевых слов, тематическом моделировании, а также будет базовым материалом в работе над изданием данных дневников.

Иллюстрации



Въ ночи въ Эрмитажѣ две тревоги;
объ не имѣвлія слѣдствій, однако же
Гря подняли во 2 м часу изъ постели и за
ставили туда ѣхать. – Сначала дымъ
показался отъ деревянныхъ опилокъ
кот. покрыты Амосовскія печи; въ
другой разъ загорѣлись какія-то трен
ки. – Случаи обыкновенные; – но что-то
непостижимое, что они именно теперь
такъ часто возобновляются. – Эена
моя была въ большой тревагѣ. – И не
у насъ однихъ пожары; въ Лондонѣ
сгорела королевская биржа; въ Парцжѣ

Пример расшифровки фрагмента 1-го Тома

Литература

1. 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, Vol. 45, P. 508 - 524