Information Extraction and Entity Linkage in Historical Crime Records: **OCR quality scoring and post-correction**

Callum Booth (cwbooth1@sheffield.ac.uk)

PRIMARY SUPERVISOR - DEPT. OF HISTORY
Professor Robert Shoemaker

SECONDARY SUPERVISOR - DEPT. OF COMPUTER SCIENCE
Professor Robert Gaizauskas

THESIS ADVISOR - DIGITAL HUMANITIES INSTITUTE
Michael Pidd

Introduction

We intend, in this project phase, to cull the British Library Newspapers (BLN) corpus down to a small working corpus of high quality and highly relevant OCR, and correct as many errors as possible automatically. In the later stages of this PhD project, we intend to use this high quality, cleaned OCR, to corroborate criminal histories held in the Digital Panopticon, using named entity recognition and relation extraction techniques, with the ultimate goal of giving historians structured access to newspaper scan text pertinent to crime and criminal lives in the nineteenth century.

Initial Corpora

We take advantage of two data sources:

- The BLN dataset—a corpus of OCR transcriptions, transcribed by Gale, of newspaper microfilm scans of nineteenth century newspapers held by the British Library—a corpus with a high number of errors, due to various factors including typeface, paper quality, historical wear and tear, degradation, microfilm quality, and OCR engine.[1]
- A subset of the Proceedings of the Old Bailey Online (OBO)—a manually rekeyed corpus of OCR transcriptions of the Proceedings of the Old Bailey.

Both datasets differ in genre and prosaic style—newspaper crime reports have the potential to reflect biases and editorialised styling in short form accounts, whereas OBO documents, while covering similar subject matters, are longer and more objective accounts of criminal trials.

We begin forming the working corpus by analysing the publication metadata of the BLN corpus. From figure 1, we see that that the corpus consists of 61 publications, only 17 of which are London-specific, allowing us to make an immediate 71.4% reduction in corpus size. We can also observe, in figure 2, that transcription counts are temporally skewed towards the end of the nineteenth century, and that some publications exist on shorter and more staggered timespans.

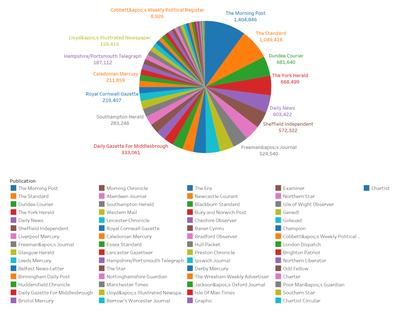

Figure 1: Document count breakdown per publication in the full corpus.

Figure 2: Document count breakdown per publication and year in the full corpus.

OCR Quality Modelling Methodology / Results

- We theorise that probabilities from a language model of a genre-adjacent corpus are valid measures of OCR quality.
- We select the Proceedings of the Old Bailey Online as the training corpus.
- The training data is split by decade and 10 separate language models are trained in order to capture the temporally staggered nature of the corpus observed in figure 2.
- The language model architecture consists of an interpolated model of bigram, unigram, and zeroth order model probabilities.[2]
- We select bigrams as the highest n-gram size to avoid capturing too much genre-specific information, but to still capture basic sentence structure.
- We score all documents after text normalisation by taking the log-sum of all interpolated bigram probabilities from the appropriate decade model.
- We assert that each model works by comparing its output score for a piece of OCR text, and a manual rekeying as a gold standard.
- We find that 7/10 of the models are correct.
- All documents in the London sub-corpus are scored, leading to the average score spectra shown in figure 3.

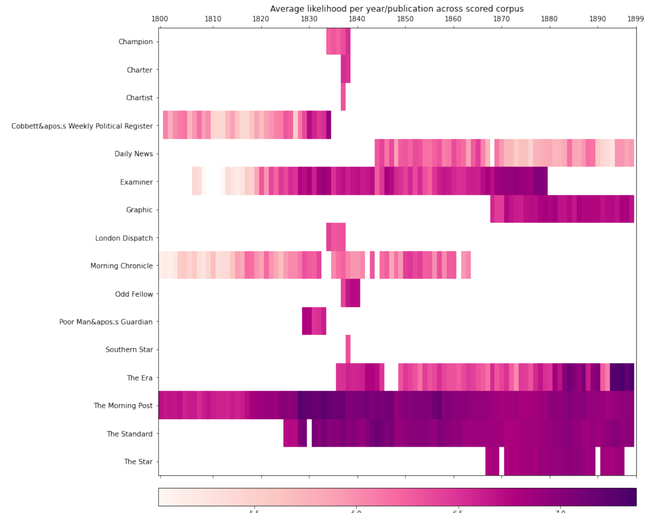

Figure 3: Average language model OCR scores per publication and year.

Corpus Culling Methodology / Results

- We divide the quality-ranked, London-specific sub-corpus into percentiles, and manually analyse the quality of the OCR in each.
- We select the 10th percentile as a compromise between OCR quality and corpus size, however other documents of reasonable quality are present in lower percentiles.

1st percentile	10th percentile	11th percentile	100th percentile
A garrison court-martial was beld on Saturday, at the Royal Artillery barracks, for the trial of several prisoners charged with insubordination and desertion.	Harry Walker, stoker, 24, of Mirfield, was indicted at Leeds Assizes yesterday for the murder of Mary Ann Chapman, whom he was alleged to have thrown over a bridge into the river at Dewsbury during a drunken quarreL	At the Liverpool Assises on Monday, George Allcock, aged 23, was indicted for hll- ing feloniously slain and killed Louisa Teel T* _. j that on the 10th of October	' Yesterdaj't Gentleniani was charged vilvh urossly in. fdulting anothar at Sadler't Wells, on ui i n;t, in consequence of Pa dispute .fdr a seat in o b);

- We observe above, the degradation of quality from the 1st percentile through to the 100th, which contains no discernible information.
- We compare the selected 10th percentile with the 11th, and find that while there are still entities discernible in the sample text, we can afford to cut the corpus at the 10th quality percentile—culminating in a corpus of approximately 338,000 documents, as opposed to the 14 million in the full corpus.

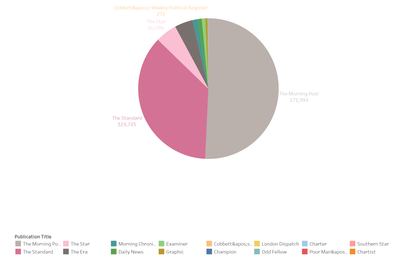

Figure 4: Document count breakdown per publication in the working corpus.

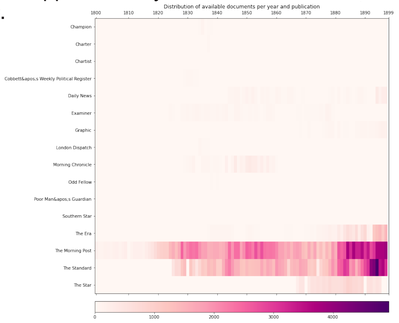

Figure 5: Document count breakdown per publication and year in the working corpus.

- Figures 4 and 5 show the final breakdown of the working, culled corpus.
- Reducing the corpus by both publication location and OCR quality reduces the corpus by a total 97.6%.
- We observe that reducing by OCR quality over the entire corpus rather than over each decade skews the corpus towards the end of the nineteenth century.
- To mitigate this, we cull within each decade in a second corpus to enforce temporal consistency.

Conclusions and OCR Post-correction

For erroneous OCR such as the BLN corpus, we were able to utilise high quality OCR from an adjacent genre to train language models that can score OCR quality. Furthermore we were able to use this model paradigm to reduce the number of documents in the BLN corpus by 97.6%, allowing us to focus our next efforts in post-correction, NER, and relation extraction, on a much higher quality dataset.

Current ongoing work on this phase of the project focuses on attempting to correct the highest quality OCR through neural methods:

- We re-frame the task as a neural machine translation task.
- We follow previous attempts at neural OCR post-correction, such as Gene Lewis, and Hämäläinen and Hengchen (2019), using architectures such as sequence to sequence encoder/decoder networks, and word2vec-aided token clustering methods.
- We attempt to utilise already existing parallel corpora, such as Trove (a rekeyed corpus of Australian newspaper OCR), ECCO, and the ICDAR2019 POCR competition dataset.
- Where high quality parallel corpora cannot be sourced for this task, we seek to generate synthetic training data, by augmenting a gold standard corpus with common OCR substitutions and omissions.

Footnotes and References

[1] The OCR engine used varied throughout the transcription process, including engines such as CCS docWorks and ABBYY FineReader. Personal communication from Tom English and Chris Houghton, October 26, 2020.

[2] The zeroth order component consists solely of 1/|V|, where |V| is the size of the vocabulary. This component acts as smoothing to deal with out-of-vocabulary errors.