Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5859–5864
Marseille, 20-25 June 2022

A Language Modelling Approach to Quality Assessment of OCR'ed Historical Text

Callum W Booth, Robert Shoemaker, Robert Gaizauskas
University of Sheffield, Sheffield, United Kingdom
{cwbooth1, r.shoemaker, r.gaizauskas}@sheffield.ac.uk

Abstract

We hypothesise and evaluate a language model-based approach for scoring the quality of OCR transcriptions in the British Library Newspapers (BLN) corpus parts 1 and 2, to identify the best quality OCR for use in further natural language processing tasks, with a wider view to link individual newspaper reports of crime in nineteenth-century London to the Digital Panopticon—a structured repository of criminal lives. We mitigate the absence of gold standard transcriptions of the BLN corpus by utilising a corpus of genre-adjacent texts that capture the common and legal parlance of nineteenth-century London—the Proceedings of the Old Bailey Online—with a view to rank the BLN transcriptions by their OCR quality.

Keywords: ocr, evaluation, language modelling, information extraction from historical text

1. Introduction

This work takes place as part of a larger project to detect and link entities from nineteenth-century historical crime records in the London press, to the Digital Panopticon (www.digitalpanopticon.org, 2022)—a repository of structured criminal 'life archives'. We target Gale's British Library Newspapers (BLN) parts 1 and 2 datasets (Gale, 2022) as the primary data source.

In order to work with these corpora tractably, we must be able to mitigate two limiting factors—dataset size, and transcription quality. Not all of the 14 million documents in the combined corpus will be relevant to the project, or of a high enough quality for downstream tasks.

OCR quality within the BLN corpus is highly variable due to factors including original scan qualities (a product of microfilm/document quality, ink and paper quality, physical degradation along the spine and areas with high hand contact, topology of the paper during scan, scan resolution, etc. (Holley, 2009)), the machine transcription process used to create the OCR (a process which varied over the course of corpus creation[1]), and modern typeface biases in OCR engines (most engines are trained to handle modern typefaces, and are lacking for historical documents (Springmann and Lüdeling, 2017)).

We wish to address both the corpus size problem and the OCR quality problem by ranking documents by their transcription quality, and then discarding documents below a quality threshold. Gale's BLN platform provides an "OCR Confidence" metric that could potentially serve as a candidate quality metric, however this would be problematic in this context for a number of reasons. The OCR confidence metric provided by the platform is an aggregation of individual character confidence values pulled directly from the OCR engine used[2]. Additionally, character level confidence does not equate to accuracy of word transcription—engine-supplied confidence metrics cannot provide confirmation that a character is correct in context—verification at the word level requires human intervention (Holley, 2009). Images, formatting, and layout elements in the initial scan image may also potentially be interpreted by the engine as text, which in turn raises the OCR confidence, despite adding no information of actual value—a prominent example of this is the transcription of borders and rules as optically-similar characters, such as `1`, `l`, `|`, and `_`. Furthermore, Gale's OCR confidence metric is not available for every item in the corpus, and, for documents with a confidence value, the metric is not available in a programmatically accessible form.

We devise, therefore—through the use of language modelling methodologies—a metric to quantify the quality of a machine transcription in the context of the BLN corpus. This work will provide a re-framing of the OCR quality analysis task as an evaluation of corpus usefulness, rather than as an intrinsic evaluation method for OCR engine research and development—by formulating an OCR quality metric to rank document quality in the British Library Newspapers corpus, without access to a gold standard. We approach the task with a historical consideration—nineteenth-century newspaper crime reporting and court reports share common genre and historical parlance. We seek to exploit this genre adjacency, by using the Proceedings of the Old Bailey Online (OBP) (Hitchcock et al., 2018) as a source of clean, machine-transcribed, human-corrected texts, from which we can judge the OCR quality of newspaper crime reports, with a view to select high quality transcriptions for downstream NLP tasks that require the cleanest entity transcriptions possible.

[1] Personal communication with Gale representatives, October 26, 2020

[2] Personal communication with Gale representatives, October 26, 2020

2. Related work

The impact of OCR quality on natural language processing tasks and pipelines in digital humanities contexts is a widely researched topic (Strange et al., 2014; Traub et al., 2015; van Strien et al., 2020). However, research into evaluation of OCR quality post-transcription is more limited. Strange et al. (2014) find that machine transcription quality of historical text suffers in modern OCR engines due to myriad factors, such as degradation of the source material, image bleed, and microfilm exposure issues that require image editing in external software to mitigate—and even given these mitigations, important entities within the text such as names were often mistranscribed repeatedly.

van Strein et al. (2020) define a Levenshtein distance-based method of evaluating an OCR transcription against a gold-standard, aligned to the original text in order to assess the impact of OCR quality on NLP tasks. They find that for named entity recognition and information retrieval tasks, the distance of the OCR from the gold standard not only has a negative impact on the task accuracy, but also causes a wider, more inconsistent range of results.

OCR quality as a function of edit distance is a frequent finding within the literature—Neudecker and Clausner (2021) cite Stephen V. Rice's (1996) approach to OCR quality calculation using a modified Levenshtein edit distance calculation, an approach adopted into software packages for OCR quality calculation such as UNLV/ISRI Analytic Tools for OCR Evaluation (Neudecker et al., 2021; Rice, 1996; Nartker et al., 2005), which has been used before contextually in historical document transcription methodologies (Springmann and Lüdeling, 2017).

We may assess OCR quality through the analysis of proportions of words appearing in a standard dictionary (van Strien et al., 2020; Strange et al., 2014). This however is not an appropriate approach to measuring historical texts where heightened focus is given to entity capture—names and locations for instance, especially short-form names common in nineteenth-century newspapers (for example, the abbreviation of "William" to "Wm.").

Machine learning approaches towards OCR quality evaluation, such as treating the task as a classification tasks, and requiring a gold-standard text for comparison (Schneider, 2021), can also be seen as ineffectual for our use-case, due to the lack of gold standard transcription.

Given these methods require either a gold standard corpus, or pre-made lists of entities and dictionaries to compute edit distances, we observe that the field is lacking solutions for measuring OCR quality from angles other than intrinsic evaluation of OCR systems. Measurement of OCR quality to rank a corpus for historical research purposes is overlooked within the literature.

3. Analysing the BLN Corpus

The BLN corpus parts 1 and 2 comprises 14,060,845 documents of varying length and significance—documents are not segmented into consistent units of information, such as pages or individual articles. Document content ranges from single character transcriptions, through single sentences, articles, columns, to transcriptions spanning half a million characters—where the initial page image and/or full page OCR output was not segmented down to article level.

Corpus documents range in length from empty documents, to 553,432 characters, with an average document length of 11,118, and median length 5,628. Document lengths are mostly distributed across a small interquantile range close to the minimum length, which may be indicative of an abundance of incomplete documents, or documents scanned via a methodology with different article segmentation logic than the documents at the upper end of the spectrum. These hypotheses are however speculative, as the process used per transcription is unknown.

Documents in BLN parts 1 and 2 cover a total of 61 publications, the vast majority are mainly regional and local publications. London-centric papers make up a relatively small proportion of the entire corpus, at only 17 publications.

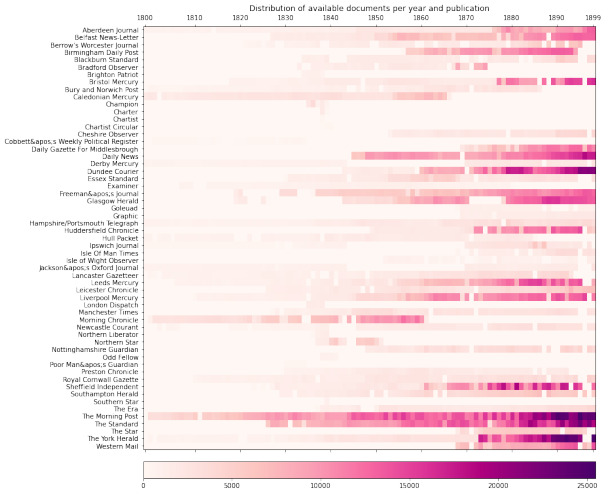

Figure 1: Heatmap breakdown of BLN transcription counts over each year and publication.

Document availability is skewed heavily towards the end of the nineteenth century—as shown by the distribution of available documents per individual publications in figure 1—which we might reason to be in part due to segmentation algorithms ran on later texts creating more individual documents versus longer, unsegmented documents, but may also be due to a higher number of publications in the late nineteenth century.

We may corroborate this in figure 2, which shows the counts of unique issues of a paper, independent of document counts. When we compare figures 1 and 2, we

Figure 2: Heatmap breakdown of BLN issue counts over each year and publication.

observe a similar skew towards the end of the century, however we do not observe a comparatively significant increase in publication counts across the space, except in a few publications. From this, we may again hypothesise the increase in documents at the end of the century to be a product of the segmentation methodology.

4. Formulating an OCR Quality Metric

We propose a language model-based approach for scoring OCR quality, by creating models that serve as semi-authoritative representations of written nineteenth-century English. Assessing OCR quality therefore becomes a matter of measuring the likelihood of the document belonging to the language model. To do this, we select a cleaner corpus of nineteenth-century prose to serve as train, test, and development sets: the Proceedings of the Old Bailey Online (OBP)—a digitised collection of the trial accounts from London's central criminal court, that claims a transcription accuracy rate of "well over 99%" (Emsley et al., 2018b).

Using the Proceedings has advantages in the scoring process—despite being of a different genre to newspaper articles—in that it provides a number of exploitable biases. The nature of its content provides a bias towards technical crime vocabulary, legal vocabulary, and crime-adjacent vocabulary used in common parlance via the inclusion of quoted testimony. As the Old Bailey was the central criminal court for London, the Proceedings also provides an additional bias towards London-centric entities.

We structure our language model as an interpolated model of decreasing dimensionality n-gram sub-models, starting at a bigram model and ending at a zeroth order. We take the zeroth order sub-model to be a Lidstone-esque uniform distribution (Hsu, 2007). By treating this as interpolation instead of backoff, the zeroth order sub-model provides Lidstone smoothing independent of the other components.

For a transcription document $\mathbf{D}$, we estimate and store sub-model probabilities using maximum likelihood estimation-based language model calculations given in equations 1 and 2, as specified in Jurafsky & Martin (2020). We estimate each sub-model's token probabilities independently, using normalised OBP tokens. Tokens are delimited by whitespace and punctuation, transformed to lowercase, and non-alphanumeric tokens are removed.

$$P_{\text{unigram}}(w_k) = \frac{\text{count}(w_k)}{\sum_{w \in \mathbf{V}} \text{count}(w)} \tag{1}$$

$$P_{\text{bigram}}(w_k|w_{k-1}) = \frac{\text{count}(w_{k-1}, w_k)}{\sum_{w \in \mathbf{V}} \text{count}(w_{k-1})} \tag{2}$$

We calculate ensemble bigram probabilities as a weighted sum of the individual sub-model probabilites in equations 1 and 2, as shown in equation 3.

$$\begin{aligned} P(w_k|w_{k-1}) = &\lambda_1 P_{\text{bigram}}(w_k|w_{k-1}) + \\ &\lambda_2 P_{\text{unigram}}(w_k) + \\ &\lambda_3 |\mathbf{V}|^{-1} \end{aligned} \tag{3}$$

where λ_i are component weights such that $\sum^i \lambda_i = 1$, that are to be conditioned on a development set after model training, and $|\mathbf{V}|$ represents the vocabulary size. We represent the OCR quality score $S_{\mathbf{D}}$ as the average log likelihood of $\mathbf{D}$, as shown by equation 4. We select this evaluation method as it provides normalisation by document length, to avoid a bias towards shorter documents.

$$S_{\mathbf{D}} = \frac{1}{K} \exp \sum_{k=1}^{K} log(P(w_k|w_{k-1})) \tag{4}$$

$$\text{where } \mathbf{D} = \{\text{<START>}, w_1 ... w_K\}$$

We opt for bigrams as the highest order n-gram supported by the model due to the genre difference between the model training material and the target OCR genre. Higher order n-grams are capable of capturing more genre-specific syntax (Jurafsky and Martin, 2020)—but we wish to capture only the local interactions between tokens, rather than overfitting to OBP syntax.

Separate model instances are trained per decade of Proceedings text to account for both linguistic and stylistic changes throughout the nineteenth century, as well as historical legal changes that alter the likelihood of the coverage of some criminal charges[3].

5. Model Evaluation and Creation of a Working Corpus

We begin by restricting the set of publications down to only London-based newspapers to align with the

[3] A number of legal changes occurred over the nineteenth century that altered the parlance used to describe certain crimes (Emsley et al., 2018a)

London-centric nature of the Old Bailey Proceedings—a total of 17 publications out of the initial 61—down to 3,388,092 documents, a 75.9% initial reduction. We carry out this initial alignment as the vast majority of criminal histories present were individuals tried by the London courts, it follows therefore that we restrict newspaper coverage to London-based newspapers to maximise the probability that a given document in the final corpus will link to the Digital Panopticon.

5.1. Model Evaluation

We sort this London-specific sub-corpus by its OCR quality score, assessed by calculating $S_{\mathbf{D}}$ of each document in the corpus against its relevant decade-specific language model. Quality sorting is performed irrespective of other data dimensions such as publication or decade, to avoid the loss of quality that would result from enforcing equal representation across the decades.

We show the lead sentences from the first documents from the top five percentiles containing court related vocabulary, when ranked by $S_{\mathbf{D}}$. For this analysis, we consider first names, last names, dates, locations, courts, and organisations as entities.

1st percentile	Gale: R3213129190
A garrison court-martial was beld on Saturday, at the Royal Artillery barracks, for the trial of several prisoners charged with insubordination and desertion.	
100% correct entities	96% correct tokens

2nd percentile	Gale: GS3214952991
The Domestic Tragedy at Glasgow.-- William Agnew, a shoemaker, was at Glasgow yesterday remitted to the High Court charged with murdering his wife, aged 40, in their house on Sunday last.	
100% correct entities	100% correct tokens

3rd percentile	Gale: R3211750099
W. M "Millar, an inmate of Prestwich Asylum, was committed to the assizes charged with the wilful murder of a fellow inmate named Jones.	
80% correct entities	96% correct tokens

4th percentile	Gale: R3214433610
John James Dewar, a boy, was brought before the Tynemoiith County Magistrates yesterday charged with causing the death of a girl, named Jane O Brady, aged 11, by shooting her with a revolver at Wal!send.	
63% correct entities	91% correct tokens

5th percentile	Gale: R3212509988
On Thursday a man of the name of William Jami- so. , who represented hinis.lfto he a farmer, was brought before the magistrate at the Central Police Court, charged with having endeavoured to impose on Peter Wallace, spirit dealer i<> High-street.	
83% correct entities	87% correct tokens

From these examples, we see that the OCR exhibits a variety of errors that make automated named entity recognition and entity linkage tasks difficult—such as corruption of names (e.g. W. M "Millar), and locations (Tynemoiith). However, it is possible to capture the vast majority of relevant entities from documents in these quality bands.

We also assess the quality of lower quality bands as a means to further check the efficacy of the quality metric:

10th percentile	Gale: R3214411435
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 nuarreL	
100% correct entities	97% correct tokens

25th percentile	Gale: R3213557389
A Miser Charged with Theft.-- At Belper en Tuesday a miser, named John .Vinson, who is report id to be worth £300, was committed for trial on a charge of stealing a spade. He has been previously convicted, acd seen s to have spent most of his time ia going about v. ith a poveitv-stiicken air begging.	
75% correct entities	85% correct tokens

50th percentile	Gale: R3213585882
Julius Adolph Deintje was indicted for obtaining on credit and under falsa pretences furniture, value £75 55., from Messrs. Carl-dan and Beaumetz, furniture dealers, of 73, Great Eastern- street, ho haviug been adjudged bankrupt.	
66% correct entities	78% correct tokens

75th percentile	Gale: BC3207264134
In the .Scond Court, before Mr. Payne, two labtouriig c men, tantied Thauoiata Spickttt and T/tantus Crisp, weore ill- di t ad for .SS.ltiog sad beating George Mauley, a Iolic- constable, in th)e eecution Of hiu duty, in GUspel Vadi k fields, ilalnipstead.	
40% correct entities	47% correct tokens

90th percentile	Gale: BB3207096330
’ Yesterdaj’t Gentleniani was charged vi!vh urossly in. fdulting another at Sadler’t Wells, on ui i n;t, in consequence of Pa dispute .fdr a seat in o b);	
0% correct entities	50% correct tokens

We may, given these additional analyses, consider the OCR metric reasonably effective, as lower rated examples progressively contain comparatively lower amounts of usable information in most cases.

5.2. Corpus Analysis

On the London-specific sub-corpus of 17 publications, we compute $S_{\mathbf{D}}$ for every document **D**, and select the top 10% as our highest quality “culled" working corpus. This culled sub-corpus contains 338,810 documents of London-specific news items, a total reduction of 97.6% from the full, whole-UK corpus. We select the 10% interval as the cut-off as it provides a sufficiently large amount of high accuracy documents to perform downstream NLP tasks. From the evaluation on the London-specific sub-corpus performed in section 5, we can see that at the 10% threshold, we still have predominantly accurate OCR, with the quality degrading at a much lower threshold.

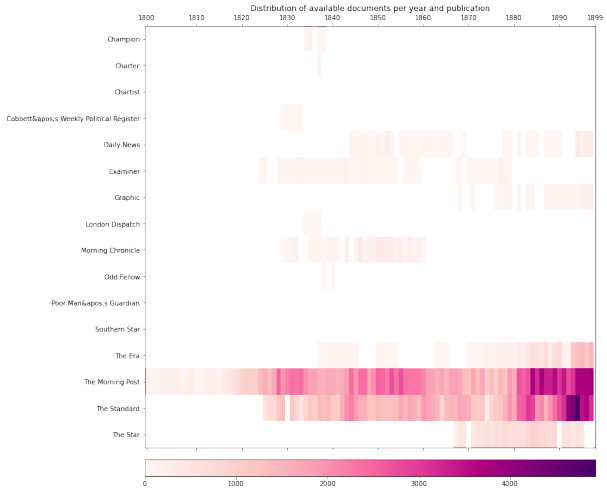

Figure 3: Heatmap breakdown of document count over each year and publication in working corpus.

The final working corpus shows significant skew towards the most document-dense publications as depicted in figure 3, with a similarly depicted skew towards the end of the nineteenth century—an expected result, as these document-rich areas will have a greater capacity to contain higher quality OCR by virtue of volume alone. In this final corpus, the bulk of our high quality data belongs to *The Morning Post* and *The Standard*—which constitute 87.2% of our working corpus, in contrast to their initial 17.4% corpus share.

Using this model and metric as a culling mechanism to remove noisy OCR has allowed us to produce a working corpus of only 3.4% of the initial BLN corpus—a still substantial corpus of 338,810 documents—that we can now use in later project efforts to perform named entity recognition, entity linkage, and relation extraction.

6. Conclusion

In this paper we have formulated a language model-based OCR transcription quality evaluation methodology by employing near-gold standard sources of adjacent genre, to assess the transcription quality of historical sources, where an existing quality metric is either inappropriate or non-existent. We assess the transcription quality of the British Library Newspapers corpus with a bias towards London-centric crime reports, by creating decade-specific language models of the Proceedings of the Old Bailey Online—a corpus of trial documentation of the Old Bailey, that provides and contextualises legal vocabulary, common parlance, and London-specific entities/locations. We apply the scoring models over the entire BLN parts 1 and 2 corpora, and use the resulting quality metric to cull the documents down to the top decile of highest quality OCR, for use in later NLP tasks. Our evaluation shows that the language model-based approach to OCR quality ranking that we have proposed is effective in identifying OCR’ed documents with a high proportion of correct entities and tokens.

7. Bibliographical References

Emsley, C., Hitchcock, T., and Shoemaker, R. (2018a). Crime and justice - crimes tried at the old bailey, old bailey proceedings online. https://www.oldbaileyonline.org/static/Crimes.jsp. version 8.0.

Emsley, C., Hitchcock, T., and Shoemaker, R. (2018b). Old bailey online - about this project, old bailey proceedings online. https://www.oldbaileyonline.org/static/Project.jsp. version 8.0.

Gale. (2022). British library newspapers. https://www.gale.com/intl/primary-sources/british-library-newspapers.

Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S., and McLaughlin, J. (2018). The old bailey proceedings online, 1674-1913. https://www.oldbaileyonline.org. version 8.0.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine: The Magazine of the Digital Library Forum*, 15(3/4).

Hsu, B.-J. (2007). Generalized linear interpolation of language models. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 136 – 140.

Jurafsky, D. and Martin, J. H., (2020). *Speech and Language Processing*, chapter 3. Stanford, third edition.

Nartker, T., Rice, S., and Lumos, S. (2005). Software tools and test data for research and testing of page-reading OCR systems. In *Document Recognition and Retrieval XII*, pages 37–47. International Society for Optics and Photonics, SPIE.

Neudecker, C., Baierer, K., Gerber, M., Christian, C., Apostolos, A., and Stefan, P., (2021). *A Survey of OCR Evaluation Tools and Metrics*, page 13–18. Association for Computing Machinery, New York, NY, USA.

Rice, S. V. (1996). *Measuring the accuracy of page-reading systems*. Ph.D. thesis, University of Nevada, Las Vegas, NV.

Schneider, P. (2021). Rerunning OCR: A machine learning approach to quality assessment and enhancement prediction.

Springmann, U. and Lüdeling, A. (2017). Ocr of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus.

Strange, C., McNamara, D., Wodak, J., and Wood, I. (2014). Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8.

Traub, M. C., van Ossenbruggen, J., and Hardman, L. (2015). Impact analysis of OCR quality on research tasks in digital archives. In Sarantos Kapidakis, et al., editors, *Research and Advanced Technology for Digital Libraries*, pages 252–263, Cham. Springer International Publishing.

van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., Mcgillivray, B., and Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, volume 1.

www.digitalpanopticon.org. (2022). The digital panopticon: Tracing london convicts in britain and australia, 1780-1925. `https://www.digitalpanopticon.org`.