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

Part of project: “Information Extraction and Entity Linkage in Historical Crime Records”

INTRODUCTION AND OUTLINE

Goal: Reduce a dataset of newspaper transcriptions down to the best quality OCR.
Project Goal: Capture details of criminal lives from nineteenth century newspaper reports, and link them to the Digital Panopticon.

Limitations:
- Gold-standard transcriptions are not available.
- OCR quality measurement is more commonly viewed as an intrinsic evaluation method for OCR systems themselves. [1, 2]
- OCR quality within the dataset is highly variable, due to various scan artefacts.

DATASETS AND INITIAL CORPUS ANALYSIS


BLN
- Transcribed by Gale from British Library microfilm.
- Parts 1 and 2 comprises over 14 million article transcriptions from nineteenth century British newspapers.

OBP
- A digitised collection of trial reports from London’s central criminal court.
- Collections of reports were regularly published whenever the court met (between 8 and 12 times a year). [6]
- Scans were transcribed through either double-rekeying, or comparison of a single-keying and a machine transcribed copy. [7]
- Transcription accuracy of “well over 99%”. [7]

MODEL ARCHITECTURE

- We fashion a language model trained on texts from the Proceedings of the Old Bailey Online.
- We exploit the genre adjacency between Old Bailey trial reports and newspaper crime reports.
- We create a weighted ensemble model of bigrams, unigrams, and a uniform zeroth order model for smoothing.
- Probabilities are estimated using MLE.

\[
P(w_k|w_{k-1}) = \lambda_1 P_{\text{bigram}}(w_k|w_{k-1}) + \lambda_2 P_{\text{unigram}}(w_k) + \lambda_3 |V|^{-1}
\]

- Separate models are trained per decade of OBP texts to account for historical changes in legal parlance. [8]

MODEL EVALUATION

- To score and rank OCR quality, we compute and sort by average log likelihood for each document in a London-specific corpus of 17 publications against its matching decade model.
- We locate the first crime report within various percentiles of the ranked dataset and manually verify the transcription quality.

FINAL DATASET ANALYSIS AND CONCLUSIONS

- The initial corpus restriction to London-specific publications represents a ~76% reduction from 14 million documents to 3.3 million.
- Using this language model methodology, we select the top 10% of documents ranked by OCR quality to form the final working corpus.
- The final corpus comprises 338k documents, a ~97% reduction from the initial corpus.
- These documents represent the best quality transcriptions of London-specific newspaper articles.
- The final corpus maintains skew towards the late nineteenth-century, and towards more document dense publications, such as The Morning Post and The Standard, which constitute 87% of the working corpus, an increase from their 17% initial corpus share.
- We conclude that we can use language modelling techniques in conjunction with adjacent genre datasets to measure and rank quality of OCR’ed historical documents.

REFERENCES


A garrison court-martial was held on Saturday, at the Royal Artillery barracks, for the trial of several prisoners charged with insubordination and desertion. 100% correct entities, 96% correct tokens