Automatic assessment of OCR quality in historical documents

A. Gupta\textsuperscript{1}, R. Gutierrez-Osuna\textsuperscript{1}, M. Christy\textsuperscript{2}, B. Capitanu\textsuperscript{3}
L. Auvil\textsuperscript{3}, L. Grumbach\textsuperscript{2}, R. Furuta\textsuperscript{1}, and L. Mandell\textsuperscript{2}

\textsuperscript{1}Computer Science and Engineering, Texas A&M University
\textsuperscript{2}Initiative for Digital Humanities, Texas A&M University
\textsuperscript{3}Illinois Informatics Institute, University of Illinois at Urbana-Champaign
What are historical documents?

- Correspondence
- Diaries
- Newspapers
- Government Documents
- Books
Digitizing historical documents

Why?
- Historical records are in analog form
- Due to their fragility, most of them are not accessible

How to make them accessible?
- Digital text transcription

Ways of digitization
- Hand transcribe each book
  - Resource intensive
- OCR: optical character recognition
  - high-error in text transcription

Mass digitization projects
Early modern OCR project (eMOP)

Goal
- Improve OCR accuracy for early modern texts
  - 300k documents, 45M pages
- Open source OCR tools

Challenges
- Early modern printing
- Document image problems
Why measure OCR quality?

Post-processing triage

*OCR Output

OCR Quality assessment

Quality Score > 0.8?

Yes → Good Documents → Linguistic Analysis

No → Bad Documents → Diagnostic Pipeline

Black, font skew, Warp, Bleed through

*hOCR output from Tesseract OCR.
Our approach

Post-process OCR output
- Page segmentation result such as bounding box (BB) coordinates
- OCR word confidence

Build ML models to remove noise
- Binary classification: classify each BB either as text or noise

Quality \downarrow \text{OCR} \propto \frac{1}{\% \text{ noise BBs}}
Language-agnostic approach
Quality assessment algorithm

OCR Output

Pre-filtering

Column segmentation

Local iterative relabeling

BB labels

Step 1

Step 2
### Prefiltering

- Provides initial labels to be refined in later stages
- Rule based classifier
  - BB properties and OCR word confidence.
  - Conjunction of rules
- Problems
  - Many text BBs classified as noise
  - Need a way to recover lost text BBs
Column extraction

– Extract individual column and then process each column
Local iterative relabeling

- Refines initial labels
- Based on BB properties and its neighborhood
- Applies an MLP classifier iteratively to refine BB labels (text/noise)

Features used during local iterative relabeling

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>Score from nearest neighbors; see eq. (1)</td>
</tr>
<tr>
<td>$C\downarrow OCT$</td>
<td>OCR word confidence*</td>
</tr>
<tr>
<td>$R$</td>
<td>Height-to-width ratio of BB*</td>
</tr>
<tr>
<td>$A$</td>
<td>Area of BB*</td>
</tr>
<tr>
<td>$H\downarrow no \ _{rm}$</td>
<td>Normalized height: $H\downarrow no \ <em>{rm} = \frac{H - H</em>{med}}{H_{IQR}}$</td>
</tr>
</tbody>
</table>

Initial labels from pre-filter stage

BBs for a column ➔ Geometric Features ➔ Multi-layer perceptron model ➔ New labels

Labels: Text or Noise

Yes ➔ No old labels = new labels?
Final output
Results

Data refinement: local iterative relabeling.

- Binarized page images
- Image
  - Multi-page; single column; ink bleed-through; multiple skew; warping; printed margins

Label creation

- Each BB returned by OCR is manually labelled as 0: noise and 1: text
- 72,366 BBs are labelled

Graph:

- X-axis: Number of iterations
- Y-axis: Proportion
- Data points:
  - 0% at iteration 1
  - 20% at iteration 2
  - 40% at iteration 3
  - 60% at iteration 4
  - 80% at iteration 5
  - 100% at iteration 5

Graph legend:

- Pre-filtering
- After iterative relabeling

EMOP corpora.
Filtering quality

- 6,775 test documents with ground truth text
- $S_{\downarrow JW}$ similarity b/w OCR output and ground truth
- eMOP uses juxta-cl* to generate $S_{\downarrow JW}$

![Graph showing change in similarity and filtering quality](https://example.com/graph.png)

*Implementations from juxtacommons.org
Discussion

Summary

- Non-text OCR outputs suffice to
  - Identify text and noise in a document image
  - Estimate the document’s overall quality
  - Improve OCR transcription performance when image processing based preprocessing is prohibitive

Future work

- Diagnostic pipeline based on active learning
- Linguistic features can be explored
Questions