The Gutenberg-HathiTrust Parallel Corpus: A Real-World Dataset for Noise Investigation in Uncorrected OCR Texts

Ming Jiang[1], Yuerong Hu[2], Glen Worthey[2][0000−0003−2785−0040], Ryan C. Dubnicek[2][0000−0001−7153−7030], Boris Capitanu[1], Deren Kudeki[2], and J. Stephen Downie[2][0000−0001−9784−5090]

[1] Illinois Informatics Institute, University of Illinois at Urbana-Champaign
[2] School of Information Sciences, University of Illinois at Urbana-Champaign
{mjiang17|yuerong2|gworthey|rdubnic2|capitanu|dkudeki|jdownie}@illinois.edu

Abstract. This paper proposes large-scale parallel corpora of English-language publications for exploring the effects of optical character recognition (OCR) errors in the scanned text of digitized library collections on various corpus-based research. We collected data from: (1) Project Gutenberg (Gutenberg) for a human-proofread clean corpus; and, (2) HathiTrust Digital Library (HathiTrust) for an uncorrected OCR-impacted corpus. Our data is parallel regarding the content. So far as we know, this is the first large-scale benchmark dataset intended to evaluate the effects of text noise in digital libraries. In total, we collected and aligned 19,049 pairs of uncorrected OCR-impacted and human-proofread books in six domains published from 1780 to 1993.

Keywords: Parallel Text Dataset · Optical Character Recognition · Digital Library · Digital Humanities · Data Curation

1 Introduction

The rapid growth of large-scale curated datasets in digital libraries (DL) has made them an essential source for computational research among various scholarly communities, especially in digital humanities (DH) and cultural analytics (CA). Particularly, recent studies in DH and CA have popularly employed state-of-the-art natural language processing (NLP) techniques for corpus analysis. For instance, scholars have been using various NLP approaches on large-scale corpora for studying culture and language evolution [1]. Meanwhile, digital libraries tend to improve their own data curation and scholarly services by advanced NLP techniques [4, 5].

While researchers enthusiastically explore the new research affordances provided by digital library and state-of-the-art NLP tools, issues concerning potential limitations of curated datasets, such as unbalanced data distribution [6], missing words [7] and OCR errors [11], have been increasingly discussed. OCR errors, in particular, are one of the most ubiquitous problems for machine-scanned digitized datasets. Despite partial exploration [9] of the effects of OCR

quality, most computational text analysis in DH and CA tacitly assumes that OCR errors do not make a substantial difference in research outcomes. Recently, more reflections and empirical investigations into the effects of OCR errors on NLP tasks such as dependency parsing and topic modeling have been undertaken, with researchers finding that some tasks could be "irredeemably harmed by OCR errors" [10].

Despite remarkable contributions made by prior investigations, we have noticed two main limitations of datasets for evaluating OCR-impacted texts and corresponding ground-truth "clean" texts. First, the paired corpora, though used for comparison, might not be exactly parallel as to their content [8]. Second, the quality of the ground-truth varies, because its generation fully relies on the crowdsourcing volunteers' edits of previously OCRed texts, without systematic examination [9, 10], leaving some doubt about the validity of OCR correction. There is scant research on large-scale paired data consisting of OCR-impacted texts and corresponding ground-truth texts that are fully proofread and/or hand-typed by humans. To fill this gap, we propose a dataset with real-world DL curated items by building parallel corpora of English-language publications from two DLs: (1) a gold-standard corpus of human-proofread and systematically checked texts from Gutenberg; and (2) a digitized corpus containing OCR errors from HathiTrust. Overall, we collected and aligned 19,049 pairs of OCRed and human-proofread books in six domains published from 1780 to 1993.

We believe this dataset could benefit a wide range of stakeholders. Digital librarians could use it to investigate potential limitations of their collections or datasets and improve digitized text quality by exploring OCR error correction techniques. Scholars utilizing DL-curated datasets for research could use this dataset to provide insights into OCR errors' impact on DH or CA research. Finally, NLP developers could study the robustness of NLP models to OCR errors by training and testing models on this dataset.

2 Data Overview

We obtained data from two well-known DLs: **1) Project Gutenberg**[3], and the **2) HathiTrust Digital Library**[4]. They were chosen for three reasons. First, both DLs provide access to full texts and metadata for a large number of volumes, thus supporting truly large-scale volume retrieval and alignment. Second, the human-proofread texts with a further systematic examination (e.g., using spell-checking, HTML validity tests, and human review) in Gutenberg make it an ideal source for building a ground-truth corpus. Finally, as one of the largest DLs in existence (with 17+ million digitized items), HathiTrust offers perhaps the largest number of overlapping data pairs, where the scanned texts in this DL contribute to building a "perfectly problematic" corpus.

To build the parallel dataset, we needed to identify the intersection between the two DLs. After a preliminary investigation, we found three primary chal-

[3] https://www.gutenberg.org/
[4] https://www.hathitrust.org/

lenges for dataset construction. First, the different nature of the corpora in Gutenberg (i.e., human-proofread) versus HathiTrust (i.e., uncorrected OCR-impacted) introduces a distinct role for duplicate items (such as reprints) in these DLs. Volume duplicates in Gutenberg are considered redundant data because the main content of duplicates is essentially the same. However, in HathiTrust, all volumes, even duplicates, contain uncertain amounts of possibly random OCR errors. Such randomness in the data makes each volume's text unique. Second, because a multi-volume work may have different structural divisions in different published versions, aligning multi-volume works between two DLs is difficult. Finally, the unbalanced and multilevel metadata (subject headings and genre labels) in Gutenberg may reduce the consistency, representativeness and differentiability of intra/interclass data. To address these issues, we present a systematic method for parallel data preparation below.

3 Method

Figure 1 shows an overall workflow containing four stages: (1) collecting and pre-processing Gutenberg metadata, (2) aligning metadata from Gutenberg to HathiTrust, (3) retrieving full texts, and (4) aligning texts.

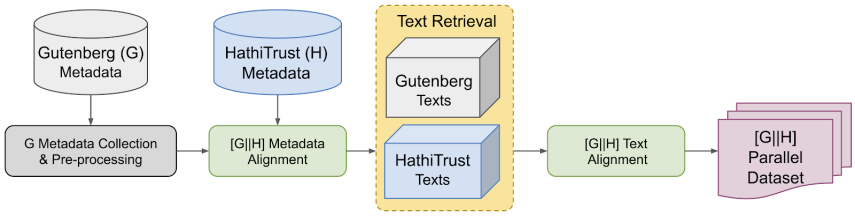

Fig. 1. Overall dataset construction workflow

3.1 Gutenberg Metadata Collection and Pre-processing

Given that any work in Gutenberg tends to have a match in HathiTrust, but not vice versa, we started with collecting and pre-processing the metadata from Gutenberg. In this study, we only considered English-language monographs. To avoid the redundancy of duplicate volumes from this DL, we randomly selected one version of each work. The number of valid matches was improved by removing any records whose author name is null, "various", or "anonymous". To obtain a set of diverse and representative works, we sampled six literature domains based on the genre labels, and took works in these domains for further analysis.

3.2 Gutenberg-to-HathiTrust Metadata Alignment

With Gutenberg metadata, we used title and author names to retrieve the metadata of matched candidates from HathiTrust. To reduce false positives,

we ruled that the query string should be exactly contained in the corresponding HathiTrust metadata records. To avoid filtering false negatives at this stage of data collection, we collected all possible matched items. To filter out any pairs containing non-English or copyright-restricted HathiTrust volumes, we double-checked their language and copyright metadata in HathiTrust.

3.3 Full-text Retrieval and Alignment

We collected texts from Gutenberg and HathiTrust based on the volume ID in each DL. To separate the front/back matter from the main content in the ground-truth corpus, we refined the data, keeping only chapterized volumes with obvious chapter landmarks like "Chapter I." in the Gutenberg repository. To align parallel texts and filter false-positive pairs, we further cleaned the paired full texts using corpus statistics, volume titles, and manual checking.

Filter by corpus statistics. Given a pair of full-text volumes (G, H), where $G = \{w_i^g\}$ and $H = \{w_j^h\}$ represents a bag of words from Gutenberg and HathiTrust volumes, we proposed three metrics:

a) the ratio of document length difference: $Num - Tks - Diff = \frac{|L_G - L_H|}{L_G}$
b) the ratio of unique tokens in H: $Num - Iso - Htks = \frac{L(H-G)_{unique}}{L(H)_{unique}}$
c) the ratio of unique tokens in G: $Num - Iso - Gtks = \frac{L(G-H)_{unique}}{L(G)_{unique}}$

L_G and $L(G)_{unique}$ denote the document length and the number of unique tokens in G. $(G - H)_{unique}$ represents the set of unique isolated tokens in G, and $L(G-H)_{unique}$ is this set size. Correspondingly, L_H, $L(H)_{unique}$, $(H-G)_{unique}$, and $L(H - G)_{unique}$ denote the similar concepts for H.

Metric a is used to identify and exclude the whole-part pairs (i.e., in which a HathiTrust volume is only a part of its paired Gutenberg text), where we set $Num - Tks - Diff \leq 0.2$. With the combination of metric b and c, we set two empirical rules to clear mismatched pairs: either $Num - Iso - Htks$ or $Num - Iso - Gtks < 0.5$, and the sum of $Num - Iso - Htks$ and $Num - Iso - Gtks < 1.2$. We set all thresholds based on our empirical investigation.

Filter by volume titles. There remained two groups of false-positive pairs caused by either a too-long HathiTrust title or a too-short Gutenberg title. The former case typically includes several subtitles, while in the latter case, the short titles often lead to ambiguities in alignment. Table 1 shows an illustrative example per group and our solutions.

Volume Title (Gutenberg)	Volume Title (HathiTrust)	Solution
Sense and Sensibility	Pride and prejudice: a novel: in two volumes/by the author of *"Sense and sensibility"*.	regular expression
Graustark	Beverly of *Graustark*/by George Barr McCutcheon.	manual check

Table 1. Examples of title-based false-positive pairs

Filter by manual checking. To further validate paired texts, we conducted a manual check on the content of paired volumes if the sum of $Num-Iso-Htks$ and $Num - Iso - Gtks \geq 1.0$. This threshold was chosen by our empirical observation of the collected data.

4 Outcomes and Analysis

4.1 Metadata Statistics

Table 2 shows the volume distribution of domains in our dataset. Overall, we collected 4,660 Gutenberg volumes and 19,049 Gutenberg-HathiTrust volume pairs. The majority of collected publications are fiction, for which there are two explanations: (1) the original literature distribution in Gutenberg is unbalanced; and (2) fiction is more likely to be chapterized than non-fiction, better satisfying our cleaning rules. Correspondingly, the most frequent subject headings are in the fiction domain.

	Fiction	Social Science	Agriculture	World War History	Medicine	Business	Total
Gutenberg	4,114	185	137	109	70	45	**4,660**
[G||H] pairs	17,060	644	487	462	185	211	**19,049**

Table 2. Volume distribution of domains

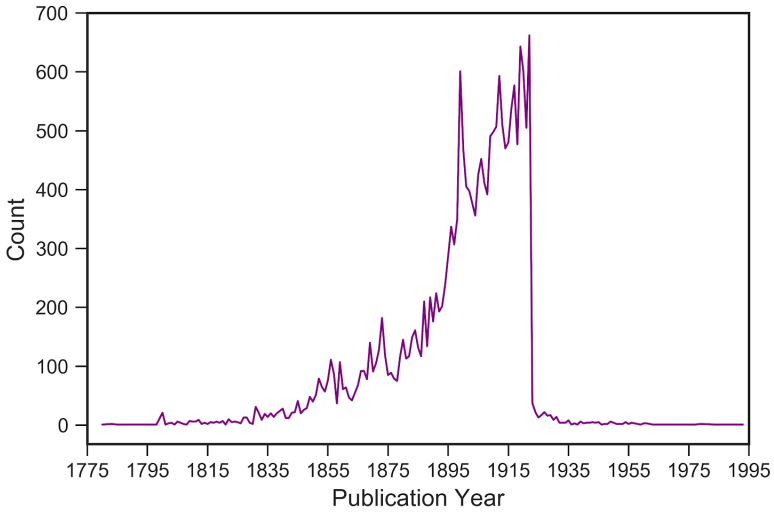

Fig. 2. Volume distribution of publication years

On average, one Gutenberg work can be matched with four copies in HathiTrust. There are 330 works having over 10 copies with various text quality.

Using the publication date provided by HathiTrust for each volume, we visualized the distribution of publication years in our dataset (see Figure 2). The

overall time-span covers 1780-1993 and our collected works have the highest intensity in the period between 1890-1922.

4.2 Corpus Statistics

To better understand the full-text characteristics, we provided corpus statistics based on the Gutenberg volume content at token-, sentence-, and chapter-levels. Table 3 provides an overview of our ground-truth corpus. In total, this corpus has over 1.2 million unique tokens, 25 million sentences, and over 130 thousand chapters. The average document length is around 110 thousand words.

	#Total Tokens	**#Unique Tokens**	**#Sentences**	**#Chapters**
Minimum	2,431	813	102	2
Maximum	670,454	26,698	35,364	365
Mean	2,431	813	102	2
Standard Deviation	60,670	2,698	3,048	18
Total	512,461,516	1,198,906	25,327,448	131,822

Table 3. Gutenberg corpus overview

	#Total Tokens	**#Unique Tokens**	**#Sentences**	**#Chapters**
Fiction	113,478 ± 60,053	8,032 ± 2,604	5,727 ± 3,000	30 ± 18
Business	110,982 ± 90,431	7,407 ± 3,769	3,860 ± 2,917	20 ± 18
Medicine	93,197 ± 74,840	7,690 ± 4,563	3,815 ± 3,846	20 ± 14
Social Science	87,576 ± 58,410	7,471 ± 3,026	3,260 ± 2,381	19 ± 13
World War History	78,189 ± 41,649	7,256 ± 2,453	3,232 ± 1,769	18 ± 10
Agriculture	68,485 ± 42,721	6,161 ± 2,781	2,686 ± 1,701	20 ± 13

Table 4. Domain-based Gutenberg corpus statistics (mean ± standard deviation)

Additionally, we did a fine-grained analysis of corpus characteristics by domain. Table 4 shows the mean and standard deviation per dimension in each domain-specific corpus. Compared to non-fiction, we observed that fiction typically had a longer volume length, larger vocabulary and more sentences and chapters. Among five non-fiction domains, the agriculture corpus is smallest by number of tokens and sentences, while works belonging to world war history typically have fewer chapters.

4.3 Case Study

Finally, to analyze the diversity of full-text quality in HathiTrust volumes, we manually examined a set of matching pairs. Two main patterns were observed in our case study. First, if the ratio of unique tokens is high in the HathiTrust volume but low in its paired Gutenberg volume, this pair is more likely to have a high amount of information noise such as OCR errors in the machine-scanned text. Figure 3 shows an example following this pattern. The isolated tokens in the HathiTrust volume are highlighted and all of them are OCR errors. Second, if the ratio of unique tokens is low, the scanned text in this pair tends to have

low information noise. For example, in the case shown in Figure 4, the volume content in HathiTrust is almost the same as it in Gutenberg.

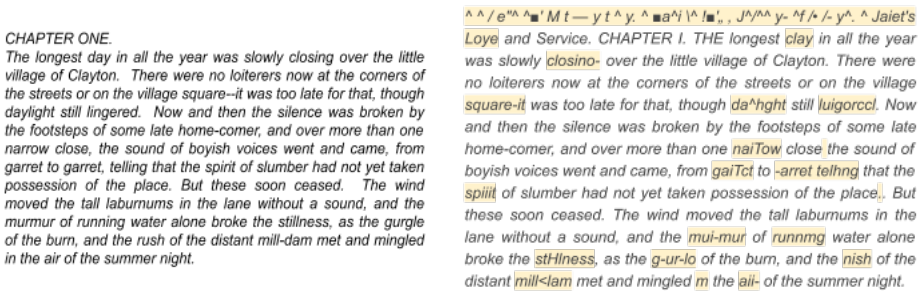

Fig. 3. A text pair example having high information noise in the HathiTrust volume

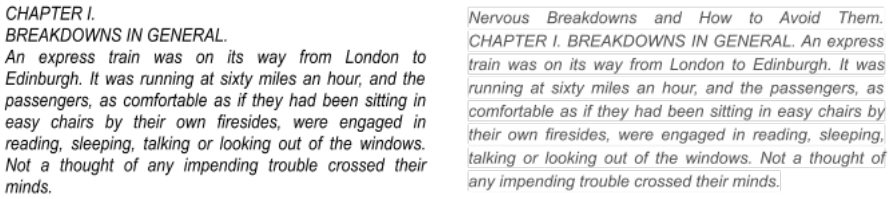

Fig. 4. A text pair example having low information noise in the HathiTrust volume

5 Conclusions and Future Work

We built a parallel dataset by retrieving human-proofread digitized texts from Gutenberg and corresponding OCR error-impacted texts from HathiTrust. This is the first large-scale benchmark dataset targeting the real-world scanned-text quality in digital libraries. With accumulated volume information from two DLs, our dataset supports fine-grained analysis along specific dimensions like publication date. Moreover, this dataset contributes to various investigations such as literary text analysis and NLP robustness. Furthermore, given chapterized texts and the large number of fiction works, our dataset benefits the exploration of advanced AI tasks such as automated storytelling.

Systematically reviewing our dataset construction process, two potential issues arise. First, the dropping of serial and non-chapterized volumes may bring biases into the data distribution (e.g., in subject domains), although these can be corrected in the future. Second, meshing different metadata schemas might

raise some minor issues such as the conflict of subject heading assignment, which can be fixed by manual review.

We plan to release the full dataset later with our upcoming study on the relative robustness of state-of-the-art word embedding models in our parallel corpus. Although all selected works are in the US public domain, general public release of text files is restricted by several DL licenses of agreement [12, 13]. Therefore, we are creating a workset to be hosted in a Data Capsule offered by HathiTrust Research Center (HTRC) for researchers to make non-consumptive use of this dataset [14, 15]. Currently, we share all the metadata and volume pairs [5] , with which full texts can be retrieved through Gutenberg API [16] and by request to HTRC.

Acknowledgements The authors would like to thank anonymous reviewers for their constructive comments on this poster. This work was carried out under the auspices of the HathiTrust Research Center, which is generously supported by HathiTrust and its member community.

References

1. Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.
2. Degaetano-Ortlieb, S., & Piper, A. (2019). The scientization of literary study. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* (pp. 18-28).
3. Underwood, T. (2019). Distant horizons: Digital evidence and literary change. University of Chicago Press.
4. Katsurai, M. (2020). Using word embeddings for library and information science research: a short survey. *ACM SIGWEB Newsletter*, (Spring), 1-7.
5. Jiang M, D'Souza J, Auer S, Downie JS. Targeting precision: A hybrid scientific relation extraction pipeline for improved scholarly knowledge organization. (2020) In *Proceedings of the Association for Information Science and Technology. 57*(1), e303.
6. Hu, Y., Jiang, M., Underwood, T., & Downie, J. S. (2020). Improving digital libraries' provision of digital humanities datasets: A case study of HTRC literature dataset. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020.* ACM, (pp. 405-408).
7. Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities, 34*(4), 825-843.
8. Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, Antoine Doucet, et al. (2019). Deep statistical analysis of OCR errors for effective post-OCR processing. In *Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries.* IEEE, (pp. 29–38).

[5] Data link: https://app.box.com/s/drlsh44qsne4p0taypqe1yzd882ldz3w

9. van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., & 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: ARTIDIGH.* (pp. 484-496).
10. Evershed, J., & Fitch, K. (2014). Correcting noisy OCR: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage.* (pp. 45-51).
11. Ryan Cordell, "'Q i-jtb the Raven': Taking Dirty OCR Seriously," Book History 20 (2017), 188-225, via http://ryancordell.org/research/qijtb-the-raven/.
12. HathiTrust Digital Library, Datasets, via https://www.hathitrust.org/datasets.
13. Project Gutenberg, Project Gutenberg Permissions, Licensing and other Common Requests, via https://www.gutenberg.org/policy/permission.html.
14. HTRC Analytics, Data Capsules, via https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule+Environment.
15. HathiTrust Digital Library, Non-Consumptive Use Research Policy, via https://www.hathitrust.org/htrc_ncup.
16. Gutenberg API, via https://github.com/c-w/gutenberg