Delving into Shakespeare's World: Cleaning Crowdsourced Transcription Data for Future Use and Reuse

ZhiCheng Wang[1] and Victoria Van Hyning[1]

[1] *University of Maryland, College Park, United States*

Abstract

This paper details our efforts in processing and organizing the dataset from *Shakespeare's World*, a crowdsourced transcription project on the Zooniverse platform. Our work focuses on addressing the usability issues of these rich but complex datasets, which encompass the crowdsourced transcription of a diverse array of Early Modern English texts from underrepresented social groups and genres. By systematically restructuring and refining the SW dataset, our work has helped created a more user-friendly resource that also serves both as a potential subject of study in the digital humanities domain and as a novel corpus for training Large Language Models.

Keywords

Crowdsourcing, Digital Humanities, Early Modern English, Large Language Models, Date Reuse

1. Introduction

Researchers in the humanities have started to adopt Machine Learning, Handwritten Text Recognition, and Large Language Models to create, enhance and explore textual corpora in new ways. However, models trained on contemporary language data are less effective when handling historical text. Previous studies have tried approaching this challenge through various methods like fine-tuning language models on historical languages [3], modernizing historic text [2] and even pre-training language models on historical text (MacBERTh) [5]. Although pre-trained models, like MacBERTh, have been shown to perform more reliably [6], the volume of text and diversity of genres of the training corpus is still limited.

Our work cleaning data outcomes from online transcription crowdsourcing projects addresses usability issues, making these resources easier to work with. By helping make the outcome of more crowdsourced transcription projects available, our work will also help enrich the training of Large Language Models (LLMs) with a unique collection of Early Modern English texts drawn from the Folger Shakespeare Library's collection. These include medicinal and food receipts (recipes) and personal letters contained in handwritten manuscripts. These genres are more likely to contain writing by women than most printed works from the fifteenth through eighteenth centuries, when women had less frequent access to formal print production [8]. In order to enhance large language models' domain-specific performance and avoid bias, we must ensure that our training data is not limited to printed works.

Over the past two decades, online transcription crowdsourcing projects have significantly contributed to the creation of digitized and searchable text from digital images of physical source materials, such as handwritten manuscripts. In 2015, Zooniverse, in collaboration with the Folger

[1]Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA

wang8769@umd.edu (Z. Wang); vvh@umd.edu (V. Van Hyning)

Shakespeare Library and the *Oxford English Dictionary*, launched Shakespeare's World (SW), featuring manuscript recipes, letters, and newsletters created around Shakespeare's lifetime [8, 9]. The project ran until 2019 and attracted 3,926 registered Zooniverse volunteers and more than 13,000 anonymous sessions (representing an unknown number of individuals). The volunteers together contributed a total of 203,390 classifications.

SW volunteers were asked to transcribe only what they felt confident they could read. Each volunteer could select the lines or words they felt confident transcribing and place a point around that segment. The points triggered a transcription box. By design, each character on each page was independently transcribed by three or more volunteers working without knowledge of what others had transcribed. The points and transcriptions were then compared utilizing a real-time aggregation method that combined a clustering algorithm step followed by a genetic sequencing algorithm (MAFFT) to compare text strings. A text string was automatically "retired" when a minimum of three people had it transcribed the same thing. Initially, the retirement rule was set to 95% accuracy [7], however in June 2016 a "hard retirement" rule was enacted to reduce the number of times a page was shown to users down from an essentially infinite number of times to 15. This was done out of concern that the clustering process was failing due to the extremely close lines of the original documents that made minor variations in dot placement were leading to aggregation failure. Our recent research reveals greater variability in data accuracy than the target 95%. Though most pages were fully transcribed at least once, fewer than 960 subjects had 80% or more of their transcribed lines achieving the targeted agree rate. The exact accuracy rates are still being determined by us and by experts at the Folger, who have ultimate stewardship for the manuscripts and their associated transcriptions. Based on manual inspection of transcriptions, and inferences drawn from prior research [1], the actual quality of what volunteers produced, and the number of pages that received sufficient transcriptions is higher than the aggregation scores suggested.

The SW data outputs, which include the full transcription and positional data provided by each volunteer, as well as expertly cleaned ground truth for several hundred pages, could provide significant data for researchers in Digital Humanities (DH), linguistics, HTR, NLP and related scholarly fields. However, this rich and complex dataset can be challenging for many teams to use, because it requires insight into the systems and processes underpinning the transcription and retirement process. The transcription interface was sunset in 2019, and the aggregation protocols are only partly documented to date [4]. This limits the exploration of deeper-level information embedded within the crowdsourcing data, like volunteer contribution patterns and the reasons behind transcriptions with low agreement scores.

Working with the raw output of the SW project (as opposed to aggregated data), we are exploring how classifications made by volunteers can be processed and organized to foster a greater understanding of the original transcription method, simplify the reuse of the project's outcome, and acknowledge its attendant challenges and affordances. By improving the reusability and accessibility of the SW dataset, our approach could serve as a case study in preparing complex, volunteer-generated crowdsourced data for reuse.

2. Cleaning the Dataset to Improve Potential Reusability

SW was the sixth transcription project on the Zooniverse platform and was highly experimental in terms of how transcription tasks were presented to volunteers and how volunteers' contributions were combined into a single reading.

When going over the dataset, we discovered multiple occasions where "`user_name`" reveals the registration emails used by the volunteers. Since an arbitrary "`user_id`" was already generated by Zooniverse for each user at registration and could be used as the unique identifier across different projects, we decided to drop the "`user_name`" field altogether to protect volunteers' privacy. Additionally, we noticed several semi-structured fields with different pieces of data stored within one attribute. For example, the original “metadata” contained the archival data provided by the Folger and the original “subject_data” contained all information captured by the Zooniverse system. To address this complexity, we broke up “metadata” and “annotation” to make the dataset flatter. In order to streamline this field, we first excluded experimental elements like “live_project” (could be easily inferred from “workflow_version”) in addition to the non-essential “user_agent” element from “metadata” field. We then separated the three remaining elements in “metadata”, “started_at”, “finished_at” and “user_language”, into independent fields.

Perhaps most significantly, we parsed out “annotation” records, which initially included all the classifications made by volunteers: identify whether the page is blank (T0), identify images or transcribe handwriting (T2) and identify if the page is completely transcribed (T3). We parsed this field into three separate fields “T0”, “T2” and “T3” according to the elements’ keys. Since the dataset was organized using “classification_id“as the unique identifier, we chose not to break up instances of multiple classifications into separate entries to avoid confusion.

The “subject_data” attribute contains metadata provided by the Folger about each manuscript document. One challenging aspect of working with this field was the variability in key names over the project's lifespan, leading to potential difficulties in both data comprehension and manipulation. To resolve this, we systematically reviewed all keys from the “subject_data” field and implemented a merge for keys that represented the same information. For instance, we consolidated keys like “Fol./p.” and “Page Number”, both indicating folio and/or page details. Additionally, recognizing the potential value for DH researchers, we meticulously isolated certain elements such as “Genre” and “Origin”, which provides crucial historical and archival context, indicating the century or year of the manuscript's origin.

3. Limitation and Future Work

We recognized that our work depends heavily on the initial structure of the SW dataset, which may limit the direct applicability of our methodology to datasets with different structures or complexities--for example other crowdsourced transcription projects from Zoonvierse or other platforms, or indeed scholarly editions of textual data. Hoping to address this concern, we plan to collaborate with other text transcription projects hosted on the Zooniverse platform and develop a data primer to assist other platform users in exploring and wrangling their own data. With a deeper understanding of different possible project outcomes, we could better facilitate different projects to transform their raw outcome into more reuse-friendly formats similar to what we achieved with the SW dataset.

We are actively finalizing the SW dataset and intend to present our work in conjunction with additional datasets relating to early modern recipe, letter and newsletter manuscripts held at the Folger Shakespeare Library in a future submission for the *Journal of Open Humanities Data.* We will make the dataset public on readily accessible platforms such as Dataverse and Github. Moving forward, we also plan to interview domain experts from DH and other related fields to gain a deeper understanding of the specific data that are most valuable for their research. This collaborative approach will not only refine our understanding of researchers' needs but also guide the development of more targeted data cleaning and structuring methodologies.

4. Conclusion

Our work with the SW dataset demonstrates one possible approach to making outcomes of online transcription projects more understandable and usable. It also reveals the inherent complexities of crowdsourcing systems that rely on independent volunteers' assessments that then get automatically compared and retired. Changes to retirement protocols or completion metrics can significantly impact data quality and thus our ability to draw inferences about volunteer performance at different points in the project. By methodically addressing these challenges, we have attempted to reduce the complexity of the "raw" data, which will, in turn, enhance our ability to distill high-quality transcriptions, and thus provide a new early modern textual dataset with reuse potential for researchers across different disciplines, including those training HTR systems, LLMs, or other models that rely on diverse and accurate textual datasets.

5. References

[1] S. Blickhan, C. Krawczyk, D. Hanson, A. Boyer, A. Simenstad, V. Van Hyning, Individual vs. Collaborative Methods of Crowdsourced Transcription, J. Data Min. & Digit. Humanit. Special Issue on Collecting,... (2019). doi:10.46298/jdmdh.5759.

[2] M. Bollmann, A Large-Scale Comparison of Historical Text Normalization Systems, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019. doi:10.18653/v1/n19-1389.

[3] X. Han, J. Eisenstein, Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2019. doi:10.18653/v1/d19-1433.

[4] G. Hines, Zooniverse Aggregation Engine Documentation, Release 0.9, 2017. URL: http://developer.zooniverse.org/_/downloads/aggregation/en/latest/pdf/.

[5] E. Manjavacas Arevalo, L. Fonteyn, MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950), in: Proceedings of the Workshop on Natural Language Processing for Digital Humanities, 2021, pp. 23–36. URL: https://aclanthology.org/2021.nlp4dh-1.4.

[6] E. Manjavacas, L. Fonteyn, Adapting vs. Pre-training Language Models for Historical Languages, J. Data Min. & Digit. Humanit. NLP4DH.Digital humanities in... (2022). doi:10.46298/jdmdh.9152.

[7] V. Van Hyning, Harnessing crowdsourcing for scholarly and GLAM purposes, Lit. Compass 16.3-4 (2019) e12507. doi:10.1111/lic3.12507.

[8] V. Van Hyning, H. Wolfe, More Content, Less Context, in: Archives, Oxford University Press, 2023, pp. 174–191. doi:10.1093/oxfordhb/9780198829324.013.0013.

[9] V. Van Hyning, H. Wolfe, P. Durkin, C. Lintott, S. Duca, R. Hutchings, P. Dingman, et al. Shakespeare's World - Zooniverse, 2015. URL: https://www.zooniverse.org/projects/zooniverse/shakespeares-world.