Combining Human and Machine Transcriptions on the Zooniverse Platform

Daniel Hanson, Andrea Simenstad
University of Minnesota - Tate Laboratory, 116 Church St SE, Minneapolis, MN 55455
{hans3724, sime0056}@umn.edu

Abstract

Transcribing handwritten documents to create fully searchable texts is an essential part of the archival process. Traditional text recognition methods, such as optical character recognition (OCR), do not work on handwritten documents due to their frequent noisiness and OCR's need for individually segmented letters. Crowdsourcing and improved machine models are two modern methods for transcribing handwritten documents.

Transcription projects on Zooniverse, a platform for crowdsourced research, generally involve three steps: 1) Volunteers identify lines of text; 2) Volunteers type out the text associated with a marked line; 3) Researchers combine raw transcription data to generate a consensus. This works well, but projects generally require 10-15 volunteer transcriptions per document to ensure accuracy and coverage, which can be time-consuming. Modern machine models for handwritten text recognition use neural networks to transcribe full lines of unsegmented text. These models have high accuracy on standard datasets (Sánchez et al., 2014), but do not generalize well (Messina and Louradour, 2015; Moysset et al., 2014). While modern techniques substantially improve our ability to collect data, humans are limited in speed and computers are limited in accuracy. Therefore, by combining human and machine classifiers we obtain the most efficient transcription system.

We created a deep neural network and pre-trained it on two publicly available datasets: the IAM Handwriting Database and the Bentham Collection at University College, London. This pre-trained model served as a baseline from which we could further train the model on new data. Using data collected from the crowdsourcing project "Anti-Slavery Manuscripts at the Boston Public Library," we re-trained the model in a pseudo-online fashion. Specifically, we took existing data, but supplied it to the model in small batches, in the same order it was collected. To test the model's predictive accuracy, we predicted each new line of text from a batch of data before training the model on that data.

After training on 90,000 lines of text, the model had an error rate of 12% on previously unseen data. This is slightly higher than other studies (Sánchez et al., 2014; Sánchez et al., 2015; Sánchez et al., 2016) which generally worked with cleaner, more curated data, potentially explaining the difference. This error rate also exceeds the 2.5% error rate achieved by volunteers when compared to experts. Nonetheless, the model performed identically to human performance in many cases, which can be used to improve transcription speed, if not accuracy.

We plan to incorporate this model into the human transcription process by showing the predicted transcriptions to volunteers as they transcribe. Much of the infrastructure already exists within Zooniverse due to the work on collaborative transcription done within the Anti-Slavery Manuscripts project. By showing volunteers the machine prediction, there are many opportunities for improving efficiency. If the computer prediction is correct, the volunteer can agree with it without retyping the whole line. If the volunteer does not agree, they can either correct it, or completely redo the transcription, ensuring high accuracy. This process will also improve model performance by allowing us to focus model training on more difficult text.[1]

[1] Portions of this work are derived from Daniel Hanson's University of Minnesota (UMN) Data Science Master's thesis. This work was funded in part by the UMN Digital Arts, Sciences, & Humanities (DASH) program, UMN College of Science & Engineering, and NSF-IIS 1619177.

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 215–216
Brussels, Belgium, Nov 1, 2018. ©2018 Association for Computational Linguistics

References

Ronaldo Messina and Jérôme Louradour. 2015. Segmentation-free handwritten Chinese text recognition with LSTM-RNN. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 171-175.

Bastien Moysset, Théodore Bluche, Maxime Knibbe, Mohamed Faouzi Benzeghiba, Ronaldo Messina, Jérôme Louradour, and Christopher Kermorvant. 2014. The A2iA Multi-lingual Text Recognition System at the Second Maurdor Evaluation. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 297–302.

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2014. ICFHR2014 Competition on Handwritten Text Recognition on Transcriptorium Datasets (HTRtS). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 785–790.

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2016. ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 630–635.

Joan Andreu Sánchez, Alejandro H. Toselli, Verónica Romero, and Enrique Vidal. 2015. ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1166–1170.